Molecular analysis of the *shaking-B* locus

of *Drosophila melanogaster*

PhD thesis

University of Glasgow, Institute of Genetics

Submitted: September, 1995

Dedication

To Dad

Acknowledgements

My thanks go to the following individuals in acknowledgement of their invaluable help and kindness. To Prof. M. Ashburner, for supplying codon usage tables, Dr Gunnar von Heijne for making available the *Top_Pred 3.2* program, Dr Luke Alphey for help with *in situ*s, Dr Mike Bate for generously giving his time and expertise to help our attempts to dye fill pupal neurons, and Dr Corey S. Goodman for sheer enthusiasm and encouragement.

I am endebted to all my friends and colleagues from the Institute of Genetics in Glasgow. I thank especially Dr John Sentry for all his help with computing, Dr Marshall Stark for advice on oligonucleotide synthesis, Dr Jerry Mottram for helping me get started with sequencing, Mary Burke for her incessant kindness and awe-inspiring competence and the expert, generous, and unfailingly kind media room ladies Doris, Agnes, Janet, Mattie, Mary, and Sadie. Thankyou Dr Sergei Korneev for regaling me with tales of Russian baths. My fellow students were ever a source of friendship, solidarity, wilderness holidays and intriguing conversation, both scientific and otherwise. I thank Martin Todman, Colin Milligan, Charlie Owen, Ruben Kok, Dani Segreto, Arvind Hundal, and Anne (née) Davidson in particular. Thanks to my colleagues in the Davies lab: Martin (again), Kevin (whichever lab he was in), Margaret, Marian, Tony, Shuqing, Helena and Mary, and Alan, for all his help and patience in getting me started.

Special thanks go to Margaret White and Linda Kerr, with heartfelt best wishes from their surrogate nephew, and to Claire Ryan, for all sorts of reasons. Thankyou Christabel for helping in the final stages of thesis preparation and for so very, very much else.

Finally, my thanks go to Dr Jane Davies. Her help and guidance made this work possible, while her kindness, support and friendship have made it a pleasure. Thankyou Jane.

Contents

Chapter Three

Chapter Four

Summary

One strategy for the isolation of molecules required for the establishment of specific synapses is to screen for mutations which disrupt identified neuronal connections and subsequently to clone and characterise the genes involved. The giant fibre system of *Drosophila melanogaster* is an ideal focus for such investigations. This system mediates the fly's jump-escape response, allowing neuronal connectivity mutants to be isolated as a subset of those flies which fail to jump in response to a light-off stimulus. The *Passover* mutation (Thomas, R. B., and R. J. Wyman, 1984. Nature **298**, 650-651) was isolated in this way, and was subsequently shown to be an allele of the *shaking-B* (*shak-B*) locus, thus implicating *shaking-B* in the establishment of neuronal connectivity, and inspiring the molecular analysis which is reported here.

Genetic analysis (Baird, D. H., A. P. Schalet and R. J. Wyman, 1990. Genetics **126**, 1045-1059) reveals two distinct functions at the *shaking-B* locus, one (termed *shak-B(neural)*) is required for the normal development of the imaginal nervous system, while the other, *shak-B(lethal)* is an essential function, without which the animals die as embryos or first instar larvae. Most *shaking-B* alleles disrupt both of these genetic functions, although some (like $shak\text{-}B^{Passover}$) are specifically neural, while others (such as $shak\text{-}B^{L41}$) affect only the essential function.

The 19E3 polytene region in which *shaking-B* resides was cloned by chromosome walking from microcloned entry points and the breakpoints of deficiency chromosomes which encroach upon the *shak-B* region were used to define a 15 kb stretch of walk in which at least some of the gene must lie. Unique DNA fragments from this area were used to probe cDNA libraries and the embryonic cDNA KE2(1.8) was isolated.

The KE2(1.8) cDNA was sequenced and found to contain no extensive regions of reading frame, though an internal 122 codon open reading frame (ORF) was implicated by computer algorithms as a likely coding region, and was found be highly homologous to the N terminus of the *Drosophila* Ogre protein and to part of the *C. elegans* Unc-7 protein, both of which are implicated in nervous system development. An asymmetric PCR strategy was used to sequence this small ORF from *shak-B* mutant chromosomes, and a 17 bp deletion which is predicted to abolish translation of the ORF was found to underlie the $shak\text{-}B^{L41}$ allele.

Due to the rarity of *shak-B* cDNAs, a library screening strategy based upon inverse PCR was devised. This technique enabled the isolation of cDNAs representing a further four *shak-B* transcript forms, while yet another two cDNAs were isolated by conventional means. Sequence analysis of these clones and of the genomic regions from which they were derived has provided a wealth of data regarding the putative products and genomic organisations of these transcripts. The SIPC8 cDNA contains an ORF of 372 residues, implying a protein of 44.4 kDa with extended homology to Ogre and Unc-7. In the neural and lethal *shak-B* alleles *shak-B*$^{R\text{-}9\text{-}29}$ and *shak-B*^{EC201} this reading frame was found to be disrupted by a mutation which introduces a stop codon in a downstream exon. This finding, together with the identification of the *shak-B*^{L41} allele suggested that Shak-B(neural) and Shak-B(lethal) proteins have unique N terminal regions but converge upon common C terminal sequences. While the SIPC8 cDNA is disrupted by lesions causing lethal alleles, the P2.4 cDNA isolated by Krishnan and colleagues (Krishnan, S. N., E. Frei, G. P. Swain and R. J. Wyman, 1993. Cell **73**, 967-977) was found to contain an ORF with a unique N terminus, and a C terminus common to that of the SIPC8 reading frame. The unique N terminus of P2.4 was found (Krishnan *et al.*, 1993) to be disrupted by lesions underlying *shak-B(neural)* mutations including *shak-B*^{Passover}, so fulfilling the criteria demanded of a *shak-B(neural)* transcript.

Shak-B proteins contain hydrophobic segments suggestive of transmembrane domains, and assessment of the likely transmembrane dispositions of all putative Shak-B proteins was carried out using optimal computer algorithms. Based on these structural predictions and on the phenotypes and expression patterns of *shaking-B* and its homologues, the possible functions of Shaking-B proteins are considered.

One

Introduction

Around 10^{11} nerve cells of approximately 1000 different types (Jessell and Kandel, 1993) together form the human brain. The positions, projections, and synaptic connections of these neurons are largely stereotyped and thus the nervous system of humans, in common with those of other animals, is the most complex of the body's organs. Such complexity is all the more remarkable because relatively small genomes are able to inform the development of hugely complex structures: there are many orders of magnitude more synapses in the nervous systems of animals than there are genes in their genomes, posing the question of how such complexity can be formed using so little information.

A current goal in developmental biology is to understand, at molecular resolution, the mechanisms by which nervous systems develop. Thus we would wish to know which molecules are active at different stages of neural development, what their individual functions are, and how, at the biophysical level, such molecules act in concert to achieve the assembly of complex, functional nervous tissues. The fruitfly *Drosophila melanogaster* is an organism well suited to such investigations (Broadie, 1994). It is amenable to extremely sophisticated genetic (Ashburner, 1989) and molecular (e.g. Gloor, *et al.*, 1991; O'Kane and Moffat, 1992) techniques, and permits also electrophysiology (e.g. Broadie and Bate, 1993c) and detailed developmental studies at the cellular level (e.g. Sink and Whittington, 1991a).

1.1 APPROACHES TO THE MOLECULAR ANALYSIS OF NERVOUS SYSTEM DEVELOPMENT IN *DROSOPHILA* AND OTHER INSECTS

To identify molecules subserving particular developmental functions, two broad strategies apply. Either promising molecules may be isolated first, and their functions subsequently investigated, or developmental functions can first be defined, and the responsible molecule(s) then sought. Both approaches have been used to great effect in *Drosophila* molecular neurobiology, each having its associated strengths and limitations.

1.1.1 The reverse genetic ('molecule to function') approach

In this approach, molecules of interest are identified either by sequence homology with those of known relevance in the development of other animals' nervous systems (e.g. Ng, 1989), or by virtue of expression patterns consistent with a role in neural development. Techniques by which molecules with interesting expression patterns may be isolated from insects include monoclonal antibody screens (e.g. Goodman, *et al.*, 1984), enhancer detection screens (e.g. Bellen, *et al.*, 1992), and subtractive hybridisation.

Having identified candidate molecules, assays of their function must then be undertaken. Powerful techniques for such investigations include the use of antibodies to attempt to perturb cellular development (e.g. Kolodkin, *et al.*, 1992), and the use of genetic techniques (e.g. Kaiser, 1990) to generate mutations in the candidate loci. Subsequent detailed phenotypic analyses of flies carrying such mutations may reveal defects in the nervous system (and possibly other tissues too), so demonstrating that the candidate molecule is genuinely active during nervous system development.

The contribution of the reverse genetic approach to the study of insect molecular neurobiology is unquestionable, yet, in a philosophical sense, such methods are not, in themselves, sufficient because the selection of molecules for study is inevitably swayed by the preconceptions of the investigator. For example, according to the *labelled pathways hypothesis* (§1.2.1.a) the outgrowth trajectories of developing axons are determined largely by specific recognition of molecular labels on the cellular substrates over which the axons grow. A large amount of distinguished work done in the laboratories of C. S. Goodman and others has resulted in the isolation of molecules such as Neuroglian (Bieber, *et al.*, 1989), Fasciclins I, II, and III and Semaphorin I (Elkins, *et al.*, 1990a; Harrelson and Goodman, 1988; Kolodkin, *et al.*, 1992; Snow, *et al.*, 1989). All of these are cell-surface glycoproteins, and each was originally isolated as the cognate antigen of a monoclonal antibody raised against membrane proteins of developing nervous tissue. Thus the question being asked was 'which molecules, by virtue of their expression patterns and cellular localisation, might mediate the recognition of labelled pathways ?', rather than 'which molecules have key functions in axon pathfinding ?' While the former question is doubtless of great interest, its scope is restricted, and questions of this style are unlikely ever to allow a comprehensive understanding of nervous

system development. In the hope of achieving such understanding, complementary approaches are required.

1.1.2 The 'function to molecule' approach of classical genetic analysis

In the classical genetic approach of first isolating a mutation which confers a phenotype (in this case a nervous system defect) and subsequently identifying the gene responsible, the molecules uncovered are less prone to investigator bias. This approach has the potential to identify genes with important functions, whose primary structures may be novel or unsuspected, and whose expression patterns may fail to evoke attention. Moreover, the existence of a mutant phenotype at the outset gives the investigator confidence that their gene(s) of interest genuinely have a relevance to neural development, whereas for some genes initially identified by molecular criteria (e.g. *connectin* (Nose, *et al.*, 1992) and *fasciclin III* (Snow, *et al.*, 1989)), such functional relevance has proven laborious to demonstrate *in vivo* (Chiba, *et al.*, 1995; Nose, *et al.*, 1994).

Conversely there are also disadvantages inherent in the 'function to molecule' approach. The principal disadvantage is that, in order to stimulate investigation, a mutation induced in a gene active in neural development must confer a detectable nervous system phenotype. There are, however, several theoretical reasons why a mutant phenotype might not be apparent. Firstly, a gene may be essential at an early stage in a non-neuronal tissue and only later be required in the nervous system. Mutagenesis of such a locus would confer early lethality without a neuronal phenotype, hence its involvement in nervous system development might go undetected. Secondly, a mutation might confer a phenotype which is too subtle to be detected by any phenotypic analysis practicable in a large-scale mutagenesis screen[1]. A subtle phenotype may result either when genes with a restricted developmental role are mutated, or when the function of the mutant gene may be complemented by one or more other loci. Such genes may be said to show *redundancy*.

It is now apparent, from reverse genetic studies in a variety of organisms, that genetic redundancy is a common and widespread phenomenon (Brookfield,

[1]The need to detect subtle morphological changes has, in some studies, been circumvented by screening for behavioural phenotypes, the rationale behind this approach being that a phenotypic change may be slight or undetectable at the morphological level, yet might lead to a measurable change in behaviour. The mutagenesis screen from which the *Passover* allele of *shaking-B* was isolated is a good example of this strategy and is discussed in detail in chapter 3.

1992; Dover, 1993; Thomas, 1993). If two or more loci were completely redundant in function then no selective pressure would exist to maintain both genes. For this reason, it has been argued that perfect redundancy can only occur for evolutionarily recent gene duplications where neither copy of the gene has yet drifted to become non-functional (Thomas, 1993). There is, however, no reason why partial redundancy of function among genes cannot be an evolutionarily stable state, as long as organisms with functional copies of the redundant genes have a selective advantage over those in which one or other redundant locus is mutated. A selection of ways in which such a selective advantage might be conferred has been mooted (Thomas, 1993).

Redundancy of function is known to be present among genes active in *Drosophila* nervous system development. *fasciclinI* (*fasI*) null mutants, for example, exhibit no gross disorganisation of the central nervous system (CNS), although adults are uncoordinated in their movements (Elkins, *et al.*, 1990b). Similarly, null mutations in the gene encoding the Abelson tyrosine kinase (*abl*) confer no visible CNS defects upon the embryo, although homozygous mutants die in the pupal period (Gertler, *et al.*, 1989). However, when flies are doubly mutant for *fasI* and *abl*, major CNS developmental defects are apparent (Elkins, *et al.*, 1990b). This form of phenotypic enhancement between null alleles at two loci is a robust indicator of functional redundancy (Thomas, 1993) and implies that the two genes have partially complementary functions, or are each components of distinct, but partially redundant pathways involved in CNS development.

Thus there exists a variety of reasons why mutations of genes active in nervous system development might not yield detectable phenotypes, and this is the major disadvantage of the classical genetic approach to the molecular dissection of nervous system development. While this drawback implies that mutational dissection will never yield an exhaustive catalogue of genes relevant to nervous system development (so emphasising the need for complementary reverse genetic approaches), any individual gene identified by virtue of its neural mutant phenotype can confidently be said to play a role in neural development. Precisely because of this confidence, it is the 'function to molecule' approach upon which the work described in this thesis is based. As discussed in chapter 3, the function in question is the establishment of a particular, well-defined, electrical synapse in the thorax of the *Drosophila melanogaster* adult. The phenotypes conferred by certain mutant alleles at the

shaking-B locus demonstrate that this gene is involved in the establishment of the synapse. However, genetic analysis of *shaking-B* demonstrates that it also has other functions, notably being required for embryonic viability, and, in order to derive as many clues as possible about what Shaking-B proteins do, considerable attention has been focussed on this vital embryonic function. Chapters 3 and 4 are devoted to unravelling the molecular structure of *shaking-B* and demonstrating how the different functions of the gene relate to this structure. In chapter 5 the structures of Shaking-B proteins and their homologues are scrutinised in detail. In chapter 6 I discuss this work, and draw together the various strands of data regarding the molecular structure of the *shaking-B* locus, optimal models of Shaking-B protein structures, and the phenotypes of the mutant alleles, to enable informed speculation as to the biochemical functions of the products of *shaking-B*. In order to establish a framework for this discussion, I here briefly review current understanding of the mechanisms by which neurons recognise and establish synapses with their appropriate target cells, and the molecules thought to be involved in these processes.

1.2 THE ESTABLISHMENT OF SPECIFIC SYNAPTIC CONNECTIONS

1.2.1 Pathway selection

The establishment of specific synaptic connections involves a hierarchy of separable developmental events (reviewed by Goodman and Shatz, 1993). Firstly, the growth cones of developing neurites must extend, sometimes for long distances, from their sites of origin on the neuronal somata towards their appropriate target regions, a process known as *pathway selection*. Early growth cones pioneer the scaffold of embryonic central axon tracts and peripheral nerves by interacting with certain glial cells, specialised midline cells and other non-neuronal substrates (e.g. Bastiani and Goodman, 1986; Klämbt, *et al.*, 1991; McConnell, *et al.*, 1989). Many studies on insects have demonstrated that such cells can act as *guideposts*: cellular stepping stones marking out trajectories along which pioneer axons travel (e.g. Bate, 1976; Bentley and Keshishian, 1983; Taghert, *et al.*, 1982). Along their routes, pioneers encounter a series of *choice points*: crossroads at which different pathway routes diverge. At such points, different growth cones select, with exquisite fidelity, pathway choices appropriate to their cell (e.g. VanVactor, *et al.*, 1993). The majority of growth cones extend later, through an environment containing preexisting axon

pathways. Such growth cones selectively adhere, or *fasciculate*, to existing axons throughout most of their routes.

1.2.1.a The labelled pathways hypothesis

Selective fasciculation has been intensively studied in the developing grasshopper CNS (reviewed by Goodman, *et al.*, 1984). At about 40% of development, the G neuron is extending its growth cone in a CNS hemisegment already comprising some 100 or so axons (Bastiani, *et al.*, 1984). These 100 axons are organised into around 25 distinct fascicles, and the profuse tufts of growth cone elaborated by the G neuron contact the surfaces of most, or all, of these fascicles. Despite the accessibility of so many axon tracts, the G growth cone invariably choses to fasciculate with just one: a bundle of four axons called the A/P fascicle. Such observations of the remarkable fidelity of pathway selection in the developing insect CNS led to the formulation of the *labelled pathways hypothesis* (Raper, *et al.*, 1983), an idea similar to earlier proposals by Ghysen and Janson (Ghysen and Janson, 1980).

The labelled pathways hypothesis states that the axon fascicles present in the embryonic neuropil in the grasshopper CNS each carry distinct cell surface molecules which serve to label them. These molecular labels are used by growth cones to distinguish the appropriate fascicle from among those lying within filopodial reach. A broadly similar situation is believed to apply to the navigation of pioneer growth cones between guidepost cells (e.g. Bastiani and Goodman, 1986; Bentley and Keshishian, 1983; VanVactor, *et al.*, 1993). Indeed, many axons which selectively fasciculate onto existing pathways will also pioneer a short part of their route, while in some cases follower axons can take on a pioneering role, albeit slowly, and with some misrouting, after ablation of the neuron which would normally pioneer their trajectory (Bentley and Keshishian, 1982; Lin, *et al.*, 1995). Thus the distinction between pioneer and follower neurons is one of degree, and the labelled pathways hypothesis has been applied to both.

1.2.1.b Pathway selection decisions suggest qualitative differences between pathway labels

Two systems of pathway labelling might be envisioned: either pathways each might have qualitatively different labels, or different pathways might be

differentiated by the presence of different amounts of common labelling molecules. If the latter system were operating, growth cones might be expected to make reproducible 'second best' pathway choices when their normal pathway was unavailable. In fact, experiments in developing insect nervous systems suggest that pathway choices generally reflect absolute preferences. In the grasshopper, for example, when the P axons are ablated, the G growth cone behaves abnormally and does not show affinity for *any* of the remaining axons. Similarly, at 30% of development, the growth cone of the anterior corner cell (aCC) is normally seen to stall for some 10-15 hours until the U1 and U2 neurons come within its filopodial range, whereupon it turns posteriorly and follows the fascicle newly pioneered by the U axons. Ablation of the U cell bodies before they extend their growth cones causes the aCC axon to travel for a very short distance anteriorly, then fail to choose another pathway from among the other four axons available at that time (du Lac, *et al.*, 1986).

In the developing neuromuscular system of *Drosophila*, mutations of the genes *beaten path* (*beat*), *short stop* (*shot*) and *stranded* (*sand*) cause growth cones of the motorneurons of segmental nerve branch b (SNb) and the intersegmental nerve (ISN) to inappropriately negotiate choice points on their routes towards their target muscles (VanVactor, *et al.*, 1993). In *beat* mutants, growth cones of the SNb stall at a choice point where they would normally separate from the ISN and segmental nerve a (SNa) motor branches, again apparently reflecting an absolute preference for a pathway that is no longer recognisable in the mutant environment.

The *Drosophila* neuromuscular system, however, also provides apparent exceptions to this 'absolute preference' rule. Mutations of the genes *short stop* (*shot*) and *stranded* (*sand*) cause growth cones of the motorneurons of segmental nerve branch b (SNb) and the intersegmental nerve (ISN) to inappropriately negotiate choice points on their routes towards their target muscles (VanVactor, *et al.*, 1993). In *sand* mutant embryos, the ISN growth cones fail to select the correct trajectory at their second choice point. At this point the growth cones are extending along afferent axons from the dorsal cluster of sensory neurons (Ghysen, *et al.*, 1986), and would normally interact with the persistent *twist* expressing guidepost cell PT3, followed by PT2 (Bate, *et al.*, 1991) and then change direction to travel past the main tracheal trunk and along the inner surface of the dorsal muscle fibres. Instead, in *sand* mutants, the ISN axons often form conspicuous contacts with the tracheal system, which apparently

acts as an acceptable alternative ISN substrate in these mutants. An analogous defect in pathway selection is apparent in *shot* mutants, in which the ISN growth cones apparently fail to respond to their encounter with the PT2 cell, and instead of extending along the inner surface of the dorsal group of muscles, probe the cell bodies of the dorsal sensory neuron cluster. One interesting potential explanation for these phenotypes is that the *sand* and *shot* mutations each disrupt the interaction between growth cones and persistent *twist* expressing guidepost cells, an interaction which would normally induce the growing axons to express different receptors for the guidance cues appropriate to the next stage of their journey. In the absence of such interactions, preferential substrates might be those along which the growth cones have been extending immediately previous to their fruitless encounters with guidepost cells (i.e. trachea until PT3 and, thereafter, sensory neurons until PT2, in the case of the ISN). Alterations in the patterns of glycoprotein expression by axons are already known to occur in response to interactions with specialised midline cells in both grasshopper and vertebrates (Bastiani, *et al.*, 1987; Dodd, *et al.*, 1988). If this explanation of *sand* and *shot* phenotypes is indeed valid, then they do not represent true exceptions to the 'absolute preference rule' and intead mimic experiments in which ablation of guidepost cells allows axonal extension to continue, but along aberrant trajectories (Bentley and Keshishian, 1983).

Although derived from results in an insect system, the labelled pathways hypothesis appears equally valid in other invertebrates and lower vertebrates (e.g. Kuwada, 1986), and may extend to higher vertebrates as well. In mammals, subplate neurons pioneer the initial axon pathways between the thalamus and the developing cerebral cortex (McConnell, *et al.*, 1989), and subsequent recognition of these inital trajectories by a labelled pathways type of mechanism has been proposed.

1.2.1.c Cell surface molecules provide short range guidance cues

The labelled pathways hypothesis has stimulated an intensive search for cell surface guidance molecules responsible for labelling axon pathways. One such search (Goodman, *et al.*, 1984) was initiated by generating monoclonal antibodies after immunising mice with crude, nervous-system enriched membrane extracts derived from 10 to 13 hour *Drosophila* embryos. The expression patterns of the cognate antigens of these monoclonal antibodies were then studied and some antibodies which stained only subsets of

fasciculating axons during axonogenesis were used for expression cloning of the antigen genes.

This approach, and other like it, have to date revealed a large number of cell surface glycoproteins proposed to play a role in axon guidance in insects (Bieber, *et al.*, 1989; Elkins, *et al.*, 1990a; Harrelson and Goodman, 1988; Kolodkin, *et al.*, 1992; Meier, *et al.*, 1993; Snow, *et al.*, 1989; reviewed by Harrelson, 1992), while a formidable number of cell adhesion molecules, integrins, and extracellular matrix components have also been proposed to have analogous functions in vertebrates (reviewed by Dodd and Jessell, 1988). For a growing number of such molecules, antibody perturbations, enzymatic degradation or genetic analyses have revealed *bona fide* functional roles in pathway selection (Burns, *et al.*, 1991; Elkins, *et al.*, 1990b; Hedgecock, *et al.*, 1990; Kolodkin, *et al.*, 1992; McIntire, *et al.*, 1992; Tang, *et al.*, 1992).

While most proteins implicated in pathway selection are proposed to mediate cell-cell adhesion, the grasshopper transmembrane protein Semaphorin I appears to act as an inhibitory cue to deter defasciculation (Kolodkin, *et al.*, 1992). The *Drosophila* protein Connectin is another cell-surface guidance molecule which has been shown to have repellent functions in pathway selection (Nose, *et al.*, 1994); it will be considered further in the discussion of target selection (§1.2.2). It is also important to note that some cell-surface guidance molecules promote different responses in different neuronal types. Thus Myelin Associated Glycoprotein, a transmembrane immunoglobulin superfamily molecule, promotes the extension of some axons while inhibiting the growth of others (McKerracher, *et al.*, 1994; Mukhopadhyay, *et al.*, 1994) and the extracellular matrix molecule Tenascin promotes outgrowth of spinal motor axons (Wehrle and Chiquet, 1990) yet provides a poor substrate for the growth of many CNS axons (Faissner and Kruse, 1990).

1.2.1.d Diffusible attractant and repellent molecules provide long-range cues in pathway selection

Selective adhesion or repulsion mediated by cell-surface guidance molecules is only one guidance mechanism influencing pathway selection in developing nervous systems. Long-range signals mediated by diffusible chemoattractant and chemorepulsant molecules are also known to be of importance (reviewed by Tessier-Lavigne, 1994). One recurring theme in central nervous system

development is that specialised midline cells provide long-range guidance signals which attract some axons to cross the midline while repelling others thus directing them to an ipsilateral route. Studies on neurons in the developing midbrain, hindbrain, and spinal cord of the mouse provide an intriguing example of such interactions. Floor plate cells in the ventral midline of the neural tube express a diffusible chemoattractant, Netrin-1, which attracts certain groups of ventrally-directed circumferential axons (Kennedy, *et al.*, 1994; Serafini, *et al.*, 1994; Shirasaki, *et al.*, 1995). Other classes of axons are known to be *deflected* from explants of floor plate (Colamarino and Tessier-Lavigne, 1995; Guthrie and Pini, 1995; Tamada, *et al.*, 1995), and recent studies demonstrate that Netrin-1, acting as a chemorepulsant may also mediate this floor plate avoidance. Colamarino and Tessier-Lavigne (1995) studied the development of the trochlear nerve of the mouse. The trochlear nerve is a wholly motor cranial nerve which arises from cell bodies lying ventrally at the hindbrain-midbrain junction (HMJ). Its axons migrate circumferentially in a dorsal direction to emerge from the HMJ at the dorsal midline, before travelling to the periphery to innervate a single contralateral eye muscle. Explanted HMJ regions reproducibly developed trochlear nerve projections around the dorsal midline, a behaviour which was completely suppressed by coculture of microdissected floor plate cells (from the HMJ ventral midline) in a collagen matrix, some 100-400μm dorsal to the explanted HMJ. Netrin-1 is expressed in the floor plate at all axial levels (Kennedy, *et al.*, 1994) and was a candidate for a molecule mediating the floor plate derived repulsion of trochlear motor axons. Transfected COS cells expressing Netrin-1 mimicked the repulsive activity of floor plate explants, while control COS cells had no effect. Thus Netrin-1 functions to repel trochlear axons, and its production by the floor plate accounts, at least in part, for the repulsion of trochlear motor axons by floor plate cells *in vivo*.

1.2.1.e *À chacun son goût*: Diffusible guidance cues may be simultaneously attractive for some axons and repulsive for others

It is striking that, as was observed for some cell-surface guidance molecules (§1.2.1.b), the diffusible agent Netrin-1 has attractant effects upon some cells and repellent effects on others. It is of interest to consider whether these different responses are due to the same receptor molecule being linked to different transduction mechanisms in different cells, or whether distinct receptors are involved. The Unc-6 protein of *C. elegans* is homologous in

structure to Netrin-1 and the similarity in apparent function of the two molecules is striking. Like Netrin-1, Unc-6 is required for circumferential migrations of some axons (and also some cells) dorsally, and for others ventrally, despite being concentrated ventrally within the worm (Colamarino and Tessier-Lavigne, 1995). The *unc-5* gene (Leung-Hagesteijn, *et al.*, 1992) encodes a putative transmembrane protein, and in *unc-5* mutants, dorsal circumferential migrations are disrupted, while ventral migrations are spared (Hedgecock, *et al.*, 1990). When Unc-5 is ectopically expressed in ventrally projecting neurons, their projections are redirected dorsally. This dorsal rerouting is dependent on Unc-6. These results together suggest that the Unc-5 protein is a receptor which recognises the Unc-6 protein as a repellent gradient. Thus, at least in some cases, different receptors are likely to stimulate different cellular responses to a particular guidance cue. Conceptually it may therefore be best to think of cell-surface molecules as signposts and chemical gradients as molecular compasses, which together provide information for pathfinding axons to interpret in a manner appropriate to each individual cell.

To date, Netrin-1 is the only identified protein with long-range chemoattractant properties for growing axons. Embryonic neurons from *Xenopus* spinal cord appear to turn up concentration gradients of the neurotransmitter acetylcholine (Zheng, *et al.*, 1994) though the *in vivo* relevance of this observation is not clear. Despite the paucity of defined chemoattractants, molecules mediating attractive signals from midline structures are implicated in many experimental systems (see Goodman and Shatz, 1993 for references). Furthermore, many experiments suggest the existence of chemoattractants derived from synaptic targets (see §1.2.2), thus it is likely that a number of chemoattractant molecules (perhaps Netrin-1 homologues) remain to be discovered.

1.2.1.f Summary of pathway selection

In summary, identified neurites faithfully select specific and reproducible pathways to their target cells. This is thought to be achieved by the growth cones of extending neurites interpreting both short range guidance cues provided by cell surface and extracellular matrix molecules, and long range cues provided by secreted molecules. Growth cone responses to guidance cues may be positive or negative, and the same molecular signal may be interpreted in different ways by different developing cells. Cellular studies suggest that growth cone decisions reflect absolute preferences for their appropriate routes,

though this has not yet been satisfactorily accounted for at the molecular level for any choice point decision.

Having considered the general mechanisms of pathway selection, the initial recognition events of synaptic target tissues will now be considered.

1.2.2 Target selection

Many parallels can be drawn between the selection of *pathways* and of *targets*, i.e. cells, or localised group of cells (neurons, muscles or glands) onto which developing axons will form initial synapses. Thus, for example, secreted, target-derived cues are likely to act in combination with cell surface target recognition molecules during target selection, through processes broadly analogous to those operating in pathway selection (§1.2.1, above). However, as discussed below, target selection molecules may be different from those with pathfinding roles (Matthes, *et al.*, 1995; VanVactor, *et al.*, 1993), and positive and negative target selection signals elicit growth cone responses distinct from those evoked by pathway selection cues.

1.2.2.a Secreted attractants and repellents in target selection

Just as long range signals act to attract growing axons to specialised midline structures, influencing their pathway selection, so too are such signals implicated in axon guidance into the synaptic target zone in a variety of different experimental systems. Thus the maxillary whisker epithelium of the mouse embryo produces long range attractant signals which guide the sensory axons of the trigeminal nerve towards their synaptic targets (Lumsden and Davies, 1983; Lumsden and Davies, 1986), while the basilar pons attracts axons which project to it from the developing cerebral cortex (Heffner, *et al.*, 1990).

While the identities of the target-derived chemoattractants implicated in these systems are, as yet, unknown, some secreted inhibitory cues involved in target selection have been elucidated and are of interest here. The development of a simple assay of growth cone collapse (Raper and Kapfhammer, 1990) enabled the purification of a collapsing protein (named Collapsin) from chick brain, and the cloning of its gene (Luo, *et al.*, 1993). The sequence of *collapsin* cDNAs revealed a high level of homology to the grasshopper *fasciclinIV* gene,

previously identified as an axonal guidance molecule required for normal pathway selection by the first tibial pioneer afferent (Ti1) in the developing grasshopper leg (Kolodkin, *et al.*, 1992). Subsequent cloning experiments have now accumulated a total of 14 homologous protein sequences from grasshopper, *Drosophila*, humans, chick, mouse and poxvirus genomes related to the original Fasciclin IV (since renamed Semaphorin I) and Collapsin sequences (Kolodkin, *et al.*, 1993; Luo, *et al.*, 1995; Messersmith, *et al.*, 1995; Puschel, *et al.*, 1995). While some semaphorin DNA sequences imply transmembrane proteins, the majority are likely to be secreted, though at least one secreted semaphorin is found to associate with cell membranes (Luo, *et al.*, 1993), hence secreted semaphorins might or might not diffuse significantly from their sites of production.

Recent studies of the functions of semaphorins have been carried out using *in vitro* experiments with chick neural tissue and genetic analysis in *Drosophila*. Semaphorin III is expressed at high levels ventrally in the chick spinal cord, but not dorsally (Luo, *et al.*, 1995; Messersmith, *et al.*, 1995; Puschel, *et al.*, 1995), and can cause local collapse of regions of growth cone, causing turning of growth cones *in vitro* (Fan and Raper, 1995), strongly suggesting an *in vivo* role in growth cone guidance through selective inhibition. Cutaneous afferent and muscle spindle afferent neurons both have their cell bodies in the dorsal root ganglion and project axons to the spinal cord via the dorsal root. However, while cutaneous afferents project to the dorsal horn only, where they synapse, muscle spindle afferents project to the ventral horn. *In vitro* studies of rat dorsal root ganglion tissue show that ventral spinal cord can repulse afferent axons, raising the possibility that a ventral inhibitory signal might be responsible for the segregation of afferent cutaneous *versus* afferent muscle spindle fibres (Fitzgerald, *et al.*, 1993). A likely role of Semaphorin III in patterning these projections has been demonstrated by recent experiments in which outgrowths of cutaneous and muscle afferent axons were independently stimulated by exploiting the different trophic requirements of the two neuron populations (Messersmith, *et al.*, 1995). Cutaneous afferent axons were deflected away from both ventral spinal cord explants and COS cells expressing *semaIII*, while muscle afferents were inhibited by neither. Thus the different responses of muscle spindle *versus* cutaneous afferents to Semaphorin III helps to account for their selection of distinct synaptic target zones.

1.2.2.b *Drosophila* Semaphorin II, a target selection cue, and Connectin, a pathfinding cue, induce different growth cone responses

In the *Drosophila* embryo, Semaphorin II, a secreted semaphorin, is expressed on a subset of neurons in the developing CNS, and by a single, large, ventral muscle fibre in the T3 segment, muscle 33 (Kolodkin, *et al.*, 1993). Its transient expression in this single muscle cell around the time of motorneuron outgrowth raised the possibility that Semaphorin II might be involved in regulating muscle target selection by motorneurons (Matthes, *et al.*, 1995). Embryos homozygous for loss of function *semaII* alleles showed no detectable neuromuscular phenotypes, though the Semaphorin II expressing muscle (33) is innervated from its ventral surface, and its innervation cannot be seen in the embryonic filet preparations used, thus a subtle phenotype restricted to muscle 33 or its motor neuron would have been missed. Matthes and colleagues next examined embryos in which expression of *semaII* was driven by the *Toll* enhancer. This caused ectopic SemaII expression in some ventral muscles which do not usually express it. Expression was strongest in muscles 28, 14-16, 7, and 6.

Ectopic Semaphorin II had no detectable effect on muscle differentiation, but caused motorneuron phenotypes, including abnormalities of the SNb projection. Normally, motorneuron RP3 (part of SNb) forms synaptic arborizations in the cleft between muscles 6 and 7, innervating both muscles. In the *Toll-semaII* transgenic embryos, RP3 axons fail to innervate muscles 6 and 7 and instead run just external to muscles 6 and 7, near to muscles 14 and 30. Although RP3 is prevented from forming terminal arborizations by ectopic Sema II expression, this appears not to repel its pathfinding, as it extends within a few microns of muscles 15, 28, 14, 7 and 6, and often ends up adjacent to muscle 14, in spite of the fact that all of these muscles express high levels of SemaII. It appears that only certain motorneurons are responsive to target inhibition by SemaII: RP1 and RP4 axons formed normal arborisations on their target (muscle 13) in the transgenic flies, despite muscle 13 expressing SemaII, albeit at lower levels.

When compared to previous studies in which the *connectin* gene was ectopically expressed under the control of the same *Toll* enhancer (Nose, *et al.*, 1994), striking differences are apparent. In the *connectin* experiments, RP3 again failed to synapse with muscles 6 and 7, but in this case, it failed to enter the

ventral muscle field in its normal postion, instead either stalling, or detouring around the *connectin* expressing muscles. Thus it is possible that SemaII inhibits target recognition by RP3, while Connectin repels its pathfinding. There is an interesting parallel to be drawn between these studies of SemaII and *in vitro* results from studies of retinal ganglion cell axons in the rat (Roskies and O'Leary, 1994). Ganglion cells from the temporal retina normally project to the rostral superior colliculus, while nasal retinal cells project to caudal regions of the superior colliculus. In a stripe assay, temporal retinal cells were found to be able to extend across alternating stripes of membrane from rostral and caudal superior colliculus. They are not repelled by the caudal membrane, but instead they preferentially branch on the correct (rostral) membrane, and do not branch on the incorrect membrane. This preference was mediated by a molecule in the caudal superior colliculus which inhibited the branching of temporal retinal axons.

The distinction between pathfinding and targeting molecules has also been vivdly drawn by Van Vactor and colleagues in their report of a saturation mutagenesis screen of the second chromosome of *Drosophila*, designed to detect mutations altering pathfinding and target selection by embryonic motorneurons (VanVactor, *et al.*, 1993). As mentioned above (§1.2.1.a), in embryos homozygous for mutant alleles of three genes (*beat*, *sand* and *shot*) growth cones of ISN and SNb axons make inappropriate pathway selection decisions at different choice points along their routes. In homozygous embryos carrying mutations at a further two loci, (*walkabout* (*wako*) and *clueless* (*clu*)), SNb motoneuron growth cones correctly navigate to the appropriate (ventral) group of muscles, yet fail to recognise their appropriate target muscle cells. This failure occurs despite the facts that growth cone filopodia can be seen to extend over muscle cells in the correct target zone and extensive investigations fail to reveal evidence of fate changes by muscle cells. Thus mutations of certain genes have no observable effects on pathway selection, but do display target selection phenotypes, emphasising that pathway and target selection are distinct processes, mediated by distinct molecules.

1.2.2.c *Drosophila* Fasciclin III may be a positive target recognition molecule

Ectopic expression in the developing neuromuscular system in *Drosophila* has also been used to demonstrate the ability of the Fasciclin III protein to function

as a positive target recognition molecule (Chiba, *et al.*, 1995). Fasciclin III (Fas III) is a transmembrane glycoprotein with three extracellular immunoglobulin domains, which functions *in vitro* as a homophilic adhesion molecule (Greeningloh, *et al.*, 1990; Snow, *et al.*, 1989). In late embryos, Fas III is expressed by both RP3 and, transiently, by its target muscles (6 and 7) in the cleft in which RP3 will subsequently synapse (Halpern, *et al.*, 1991). Fas III is also expressed by a handful of other motorneurons (such as RP1) and muscles (15 and 16). When Fas III expression was driven ectopically in all muscle cells using a *myosin heavy chain* gene promoter, RP3 was often found to innervate incorrect muscles, either instead of, or in addition to, its normal targets, implying that Fas III may contribute to target recognition by RP3 *in vivo*. Inappropriate innervation occurred in the context of a neuromuscular system whose gross morphology was indistinguishable from normal. Innervation was restricted to muscle cells in the ventral muscle field normally contacted by the RP3 growth cone filopodia (6, 7, 13, 14, 15, 16, 30), suggesting that in target selection by *Drosophila* motorneurons, groups of target muscles share a common identity, yet also have specific markers (perhaps including Fas III) which permit recognition of individual cells by motorneuron growth cones.

Ectopic muscle expression of Fasciclin III is insufficient to redirect target choice by all Fas III expressing motorneurons, as the RP1 neuron faithfully selects its normal target in spite of ectopic Fas III expression by surrounding muscle cells. This provides another example of the different responses of different neurons to a given guidance cue.

Fasciclin III is certainly not, however, the only molecule targeting RP3 to muscles 6 and 7. Target choice by RP3 in *fasIII* null mutants is normal (Keshishian, *et al.*, 1993), demonstrating that the molecule is dispensable, its function perhaps being complemented by other, partially redundant, genes. Furthermore, the expression of Fas III in *wako* and *clu* mutants is unaltered (VanVactor, *et al.*, 1993), yet RP3 fails to make its synapse onto muscles 6 and 7 in these flies. Thus muscle expression of Fas III is neither necessary, nor sufficient to direct target recognition by RP3, though the ectopic expression experiments imply that it can contribute to this process in the context of other target recognition signals.

1.2.2.d Pathway and target selection require neither neural activity nor chemical neurotransmission

Multiple lines of evidence indicate that neither pathway, nor target selection requires electrical activity in the developing neurons. Thus, for example, in the *Drosophila* embryonic neuromuscular system, the RP3 neuron forms its normal terminal arbor, on muscles 6 and 7, in the complete absence of neural electrical activity (Broadie, *et al.*, 1993; Keshishian, *et al.*, 1993). Classic experiments by V. C. Twitty and colleagues (Twitty, 1937; Twitty and Elliot, 1934) demonstrate strikingly that neuronal pathfinding, (and probably target selection) also occur normally in the absence of electrical activity in a lower vertebrate. These workers discovered that embryos of the newt *Taricha torosa* contained a toxin, tetrodotoxin (TTX), which would block neural activity in other amphibians while having no effect on *Taricha* itself. When pieces of Taricha embryos were grafted to embryos of the salamander *Ambystomid urodeles*, the host embryos were completely paralysed for several days. The toxin eventually wore off, at about the time the larvae would normally begin to feed independently. Remarkably, the experimental larvae soon began to swim and feed in a relatively normal fashion, implying that motorneuron pathways, sensory pathways and central circuits generating motor patterns, had all developed normally in the absence of neural electrical activity.

Further studies in a variety of organisms demonstrate that pathfinding and target selection probably do not require chemical neurotransmission either. In *Drosophila*, neuromuscular transmission is glutamatergic (though certain synaptic arbors may also coexpress other transmitters, such as proctolin (Anderson, *et al.*, 1988) and octopamine (Halpern, *et al.*, 1988). Argiotoxin venom from the orb weaving spider contains glutamate receptor toxins which block neuromuscular transmission. Injection of argiotoxin into embryos had no effect on axon pathfinding or target selection (Keshishian, *et al.*, 1993). A similar result emerges from studies of zebrafish *nic-1* mutants. In zebrafish and other vertebrates, neuromuscular transmission is mediated by acetylcholine. In *nic-1* mutants, the function of the muscle acetylcholine receptor is blocked (Westerfield, *et al.*, 1990), yet the observed patterns of motorneuron pathfinding and muscle target selection are indistinguishable from those of wild-type fish.

1.2.2.e Summary of target selection

In summary, secreted and cell-surface cues are are believed to act together in instructing neurons to select their appropriate synaptic targets. As was

observed in pathway selection, different neurons may respond in different ways to the same signal (e.g. the differential responses of cutaneous versus muscle sensory afferents to Semaphorin III, (§ 1.2.2.a)). While there are many parallels between mechanisms of pathway and target selection, distinct molecules are believed to mediate these different functions, and neurons respond differently to target selection, *versus* pathway selection cues (§1.2.2.b). Both pathway and target selection by neurons are believed to be independent of both neural electrical activity and chemical neurotransmission (§1.2.2.d).

1.2.3 Address selection

The degree of synaptic specificity that is achieved during target selection varies greatly among the different developmental systems in which the establishment of synapses has been studied. Thus in the *Drosophila* neuromuscular system, accurate projections between individual muscles and neurons are established by activity independent target selection processes, activity dependent processes accounting only for subsequent synaptic maturation (Keshishian, *et al.*, 1993). On the other hand, in mammalian muscles target selection processes result in 2 or more motor axons innervating each muscle fibre prior to birth, and activity dependent address selection during postnatal development leads to elimination of some of these synapses, while activity dependent address selection is even more strikingly observed in the development of the vertebrate visual system (see Goodman and Shatz, 1993 for review). These differences illustrate the emerging principle that all nervous systems use comparable pathway selection, target selection, and address selection mechanisms, but to different degrees, to establish synaptic connectivity.

While a detailed account of activity dependent address selection in vertebrates is beyond the scope of this introduction, the activity dependent maturation of *Drosophila* synapses is of potential relevance here, and is discussed below.

1.2.3.a Activity dependent synaptic maturation in the *Drosophila* neuromuscular system

During *Drosophila* embryogenesis, the processes of address and target selection establish the stereotyped pattern of neuromuscular synapses, as discussed above. The synapses are, however, immature both in morphology and in

function, and will undergo further development during late embryogenesis and larval life.

1.2.3.a.i Morphological considerations

The motor endings of the earliest *Drosophila* neuromuscular synapses are small and bear few boutons (Broadie and Bate, 1993c; Halpern, *et al.*, 1991; Johansen, *et al.*, 1989; Sink and Whittington, 1991b), and extensive growth and elaboration is found to occur during postembryonic life, with muscle fibres increasing in length by nearly ten-fold, synaptic branching order increasing four-fold, and bouton number increasing thirty-fold (Keshishian, *et al.*, 1993). Electrical activity is believed to play a pivotal role in this synaptic maturation. Thus, when the sodium channel blocker TTX is injected into stage 17 embryos (Keshishian, *et al.*, 1993), the motor ending branches on muscle fibres 12 and 13 are found to increase in number over those in wild-type controls. As no TTX-sensitive sodium channels exist in embryonic or larval muscle (Broadie and Bate, 1993c), the effect on branching is exerted, directly or indirectly by blockade of neural electrical activity. Glutamatergic neuromuscular blocking is unlikely to be responsible, as the glutamate receptor blocker argiotoxin alone has no effect (Keshishian, *et al.*, 1993).

A number of other studies also emphasise the role of electrical activity in the maturation of *Drosophila* neuromuscular synapses. Thus the extent and complexity of motorneuron branching has been found to increase in hyperexcitable larvae carrying mutations of the *Shaker* I_{KA} potassium channel (Budnik, *et al.*, 1990; Jia, *et al.*, 1993; Zhong, *et al.*, 1992), while in mutants with reduced electrical activity (e.g. nap^{ts} (Jarecki and Keshishian, 1993; Wu, *et al.*, 1978)), ectopically placed inputs are observed on muscles from motorneurons which normally innervate adjacent muscle cells. This latter situation mimics the effect of denervating specific muscle cells by laser ablating motorneurons (Keshishian, *et al.*, 1993). These observations may reflect a number of distinct activity dependent processes, and, while the mechanistic details are not yet clear, the contention that electrical activity profoundly influences morphological maturation of *Drosophila* neuromuscular synapses appears unequivocal.

1.2.3.a.ii Functional considerations

Functional maturation of the synapse between the RP3 motorneuron and muscle 6 has been extensively studied by Broadie and Bate (Broadie and Bate, 1993c; Broadie and Bate, 1993d; Broadie and Bate, 1993a). The development of synaptic function was found to occur in the following sequence: (i) motor axon filopodia explore the cell surface and begin to express neurotransmitter; (ii) myotubes become electrically uncoupled from their neighbours (see also Broadie and Bate, 1993b); (iii) the uncoupled myotubes begin to express glutamate receptors, diffusely, and at low levels over their entire surfaces; (iv) release of transmitter from motor axon filopodia is detectable, while nerve stimulation evokes excitatory junctional currents (EJCs) carried by 1-10 glutamate receptors on the muscle cell surface; (v) motor axon filopodia become localised to the mature synaptic zone; (vi) glutamate receptors are clustered at the mature synaptic site; (vii) vigorous neuromuscular activity, characteristic of larval locomotion begins and (viii) a second stage of receptor expression (at much higher levels) begins, and continues throughout the remainder of embryogenesis.

Having established this sequence in wild-type animals, Broadie and Bate addressed the question of whether the presynaptic cell acts to induce postsynaptic specialisations (i.e. glutamate receptor expression and clustering). In the *prospero* mutant (Doe, *et al.*, 1991) embryonic muscle innervation is delayed or removed completely. In the absence of innervation, muscle 6 still expresses Fasciclin III correctly, in the cleft where it abuts muscle 7 (see §1.2.2.c), and therefore must, in some sense, still be able to define the correct synaptic zone in the absence of instructive signals from RP3. However, in the absence of RP3, the normal clustering of glutamate receptors in the synaptic zone does not occur, and the second, high-level phase of glutamate receptor expression is also absent (Broadie and Bate, 1993d).

Broadie and Bate went on to demonstrate, using a variety of pharmacological techniques and mutant lines that a crucial element in the signalling by RP3 is (once again!) neural electrical activity (Broadie and Bate, 1993a). Thus, for example, in animals carrying temperature-sensitive alleles of the sodium channel locus *paralytic* (Loughney, *et al.*, 1989), the glutamate receptor field at the RP3-muscle 6 synapse develops normally at permissive temperatures, while at non-permissive temperatures, the mature receptor field fails to develop. Instead postsynaptic function (as assayed with glutamate iontophoresis) resembles that of an RP3-deprived muscle 6.

1.2.3.b Summary of address selection

In many vertebrate systems, activity dependent mechanisms refine initial patterns of synaptic projections which are established by pathway and address selection processes (reviewed by Goodman and Shatz, 1993). In the *Drosophila* neuromuscular system, activity independent events establish the pattern of neuron to muscle projections but these initial, immature contacts undergo subsequent activity dependent morphological and functional maturation. Exactly how neural electrical activity exerts these effects is not yet clear.

Two

Materials and Methods

2.1 ENZYMES AND CHEMICALS

2.1.1 Enzymes

Restriction endonucleases were obtained from Promega, Gibco-BRL and Pharmacia. Calf intestinal alkaline phosphatase (CIP), and Pfu polymerase™ were obtained from Stratagene. Sequenase™ version 2.0 (United States Biochemicals) was used for all sequencing reactions. Other DNA modification enzymes were from Promega, Boehringer Mannheim, Gibco-BRL, and Stratagene. Proteinase K, RNAase A, and lysozyme were obtained from Sigma.

2.1.2 Chemicals and membranes

General laboratory chemicals were obtained from Fisons, BDH, Sigma, May and Baker, Oxoid, Aldrich Chemical Co. and Bio-Rad Laboratories. All radiochemicals were purchased from NEN. Caesium chloride was obtained from Rose Chemicals. Acrylamide, bisacrylamide, TEMED, and ammonium persulphate were obtained from Bio-Rad. Agarose, and low melting temperature agarose were obtained from Gibco-BRL. All reagents for oligodeoxyribonucleotide synthesis were obtained from Applied Biosystems and from Cruachem. T3 and T7 promoter oligodeoxyribonucleotides were purchased from Promega. Deoxyribonucleoside 5' trisphosphates and random hexanucleotides were obtained from Pharmacia. DNA sequencing reactions were carried out using the Buffer Kit for Sequencing with Sequenase™ (United States Biochemicals).

3MM filter paper was obtained from Whatman. Southern blot and colony lift hybridisations were carried out using Hybond™ C-Extra reinforced nitrocellulose membranes (Amersham).

2.1.3 Growth media, antibiotics and indicators

Yeast extract and Bacto-tryptone for *E. coli* growth media were supplied by Oxoid. Ampicillin, tetracycline, and kanamycin were obtained from Sigma, as was the *lac* operon inducer IPTG. X-gal was obtained from Boehringer Mannheim.

2.2 ORGANISMS AND GROWTH CONDITIONS

2.2.1 Table of bacterial strains

Strain	Genotype	Source/Reference
XL1-Blue	*sup*E44, *hsd*R17, *rec*A1, *end*A1, *gyr*A46, *thi*, *rel*A1, *lac*$^-$, F'[*pro*AB$^+$, *lac*I^q, *lac*ZΔM15, Tn10(*tet*^{r})]	Bullock, *et al.*, 1987
DS941	*rec*F143, *pro*A7, *str*31, *thr*1, *leu*6, *tsx*33, *mt*12, *his*4, *arg*E3, *lac*Y$^+$, *lac*ZΔM15, *lac*I^q, *gal*K2, *ara*14, *sup*E44, *xy*15.	David J. Sherratt

2.2.2 Table of helper phages

Strain	Source	Reference
VCSM13	Stratagene	Stratagene product literature
R408	Stratagene	Stratagene product literature

2.2.3 *Drosophila* strains

2.2.3.a Table of wild type *Drosophila* strains

Strain	Source
Canton-S	University of Glasgow Institute of Genetics
Oregon-R	University of Glasgow Institute of Genetics
Qa	*Drosophila* Stock Centre, Umeå
M56i	*Drosophila* Stock Centre, Umeå
Drosophila simulans	Prof M. Ashburner, University of Cambridge

2.2.3.b Table of multilocus rearrangements

Rearrangement	Cytology	Genetic Region Deficient or Duplicated	Reference
Df(1)16-3-35	19D2,3;19E3,4	*mal* to *shak-B*	Yamamoto and Miklos, 1987
Df(1)LB6	19E4;20A2,3	*shak-B to eo*	Schalet and Lefevre, 1976
Df(1)A118	19E4,5;19E8	*R-9-28* to *vao*	Schalet and Lefevre, 1976
Df(1)17-351	nd[1]	*R-9-28 to LB20*	Perrimon, *et al.*, 1989
Df(1)HC279	nd	*R-9-28* to *vao*	G.Lefevre (unpublished); Perrimon, *et al.*, 1989
Df(1)LB7	nd	*R-9-28* to *su(f)*	A. Schalet (unpublished); Perrimon, *et al.*, 1989
Df(1)T2-14A	19E5;19E7,8	*EC235* to *vao*	Schalet and Lefevre, 1976
Df(1)17-489	19E4,5;20E	*EC235* to *bb*	Miklos, *et al.*, 1987; Lindsley and Zimm, 1992
Dp(1;Y)y+Ymal^{171}	nd	*shak-B* to *su(f)*	Lifschytz and Yakobovitz, 1978

1 nd: not determined. In every case, the approximate cytological extents can be inferred from the genetic extents.

2.2.3.c Table of single locus mutations

Locus	Allele	Mutagen	Reference
runt:	*HM449*	Hycanthone methane sulphonate (HMS)	Kramers, *et al.*, 1983
shaking-B:	*shak-B*^{2}	Ethyl methane sulphonate (EMS)	Homyk, *et al.*, 1980
	shak-B^{Passover}	EMS	Thomas and Wyman, 1984
	shak-B^{R-9-29}	EMS	Lifschytz and Falk, 1969
	shak-B^{E81}	EMS	Lifschytz and Falk, 1969
	shak-B^{EC201}	EMS	George Lefevre, (unpublished); Baird, *et al.*, 1990
	shak-B^{HM437}	HMS	Kramers, *et al.*, 1983
	shak-B$^{17-360}$	Neutrons	Baird, *et al.*, 1990
	shak-B^{EF535} [1]	EMS	Baird, *et al.*, 1990
	shak-B^{L41}	X-ray	Lefevre, 1981; Baird, *et al.*, 1990
R-9-28:	*R-9-28*	EMS	Lifschytz and Falk, 1969
EC235:	*EC235*	EMS	George Lefevre, (unpublished); Perrimon, *et al.*, 1989

1 *EF535* allele is no longer extant: see Chapter 3.

2.2.4 Culture and storage of *E. coli* cells

The following media were used for the growth and storage of *E. coli*. *E. coli* growth media were sterilised by autoclaving at 120°C, 120lb/in^2, for 20 mins. Where required, antibiotics were added to a final concentration of 100µg/ml (ampicillin), 12.5µg/ml (tetracycline) or 50µg/ml (kanamycin). Antibiotic stock solutions were made up at 1000 times their working concentrations and were stored in small aliquots at -20°C. In cloning experiments where the *lacZ* gene was used to differentiate between recombinant and nonrecombinant clones, petri dishes containing about 25ml set L. agar were dried for 30 mins at 37°C and spread with 100µl IPTG stock solution (100mM) and 20µl X-gal stock solution (4% (w/v) in dimethylformamide (DMF)). These stock solutions were stored at -20°C.

L. broth:	1% (w/v) Bacto-tryptone, 0.5% (w/v) yeast extract, 0.5% (w/v) NaCl, pH 7.0 (with NaOH)
Superbroth:	3.5% (w/v) Bacto-tryptone, 2% (w/v) yeast extract, 0.5% (w/v) NaCl, pH 7.5 (with NaOH)
2xYT:	1.6% (w/v) Bacto-tryptone, 1% (w/v) yeast extract, 0.5% (w/v) NaCl, pH 7.0 (with NaOH)
SOB:	2% (w/v) Bacto-tryptone, 0.5% (w/v) yeast extract, 10 mM NaCl, 2.5 mM KCl, pH 7.0 (with NaOH). Just before use, sterile $MgCl_2$ was added to a final concentration of 10mM
SOC:	SOB, 20mM glucose added just prior to use
ψ broth:	2% (w/v) Bacto-tryptone, 0.5% (w/v) yeast extract, 20mM $MgSO_4$, 10mM NaCl, 5mM KCl, pH 7.5 (with NaOH)
L. agar:	As L. broth, but with 2% (w/v) Bacto-Agar added prior to autoclaving
Glycerol-peptone:	40% (v/v) glycerol, 2% (w/v) peptone (Difco)

Long term storage of *E. coli* cells was achieved by adding 0.8 mls of broth culture to 1 ml of sterile glycerol-peptone and freezing at -20°C or -70°C. Such stocks could be stored indefinitely. Plate cultures could be stored for short periods (<3weeks) at 4°C. Plate cultures were grown overnight at 37°C. Liquid cultures were grown at 37°C with vigorous shaking. Specific culture procedures are described with the protocols to which they relate.

2.2.5 Titering of helper phages

Phage buffer:	1.8% (w/v) $Na_2HPO_4.12H_2O$, 0.3% (w/v) KH_2PO_4, 0.5% (w/v) NaCl, 1mM $MgSO_4$, 0.1mM $CaCl_2$, 0.001% (w/v) gelatin
Top agar:	1x L. broth (§2.2.4), 10mM $MgSO_4$, 0.6% (w/v) Bacto-Agar

100ml L. broth containing tetracycline were inoculated with XL1-Blue cells and incubated with shaking at 37°C until the optical density at 600nm (OD_{600}) reached 0.5. The cells were harvested by spinning at 3840g for 5 mins in a JA-14 rotor (Beckman) and were resuspended in 5ml phage buffer. Serial dilutions of phage (10^{-1} to 10^{-11}) were set up in 1ml volumes of phage buffer. For each dilution, 4ml top agar at 50°C were added to a mixture of 100µl plating cells and 800µl phage, then plated on prewarmed (37°C) L. agar plates containing tetracycline. Plates were incubated overnight at 37°C and the phage titre was estimated by counting plaques in dilutions at which discrete plaques were visible, and multiplying by (dilution factor x 10/8).

2.2.6 Care and maintenance of *Drosophila* stocks

Yeast-Glucose medium:	1% (w/v) Bacto-Agar, 1.5% (w/v) sucrose, 3% (w/v) glucose, 3.5 % (w/v) active dried yeast, 1.5% (w/v) maize meal, 1% (w/v) wheat germ, 3% (w/v) treacle, 1 tbsp/l soya flour; simmered for 20 mins, then supplemented with 0.5% (v/v) propionic acid and 0.1% (w/v) Nipagin M once cooled to below 70°C

Flies were grown in yeast-glucose medium in 20ml plastic vials and half pint bottles, with cotton wool bungs, at 18°C and 25°C. Stocks were shaken weekly into fresh vials. When amplifying flies in bottles, the yeast glucose food was supplemented with around 0.5ml of fresh yeast paste, and a filter paper or tissue was introduced to provide extra dry surface on which the larvae could pupariate. All sorting of flies was carried out under CO_2 anaesthesia using a paintbrush.

2.3 DNA AND cDNA CLONES AND SUBCLONES

2.3.1 Table of Genomic DNA and cDNA libraries

Library	Description	Reference/Source
Kauvar 6-12 hr cDNA	*Drosophila* 6-12 hr embryonic cDNA library constructed in λgt10	Poole, *et al.*, 1985
Nick Brown 12-24 hr cDNA	*Drosophila* 12-24 hr embryonic cDNA library constructed in pNB40	Brown and Kafatos, 1988
Stratagene adult cDNA	*Drosophila* adult cDNA library constructed in λZAP™	Stratagene Short, *et al.*, 1988
Oregon-R genomic	*Drosophila* genomic library constructed in λ EMBL4	Vince Pirotta
Canton-S genomic	*Drosophila* genomic library constructed in λ EMBL4	Vince Pirotta
dp, cl, cn, bw genomic	*Drosophila* genomic library constructed in λ EMBL3	Bill Gelbart

2.3.2 Table of λ phages and their sources

Phage clone	Insert size	Library	Isolated by
λ9405	17.3 kb	Oregon-R genomic library	J. A. Davies
λ9403	13.0 kb	Oregon-R genomic library	J. A. Davies
λ940	14.4 kb	Oregon-R genomic library	J. A. Davies
λ94C11	16.1 kb	Canton-S genomic library	J. A. Davies
λ94C15	14.5 kb	Canton-S genomic library	J. A. Davies
λ95204C9	17.1 kb	Canton-S genomic library	J. A. Davies
λ95208C9	16.9 kb	Canton-S genomic library	J. A. Davies
λKE2	1.8 kb	Kauvar 6-12 hr embryonic cDNA library	J. A. Davies
λP1	3.9 kb	Stratagene adult cDNA library	Helena Yang
λM23	3.5 kb	Stratagene adult cDNA library	Helena Yang
λM11B	2.4 kb	Stratagene adult cDNA library	Helena Yang
λM13	1.5 kb	Stratagene adult cDNA library	Helena Yang
λM12A	2.45 kb	Stratagene adult cDNA library	Helena Yang
λM14	2.9 kb	Stratagene adult cDNA library	Helena Yang
λS3A	3.4 kb	Stratagene adult cDNA library	Shuqing Ji

2.3.3 Table of plasmid clones and their sources

Name	Vector	Insert size	Description	Source / Reference
pBluescriptII™ KS+ KS- SK+ SK-			Ap^r cloning vector derived from pUC. SK and KS refer to the orientation of the polylinker (a Sac I-Kpn I fragment) + and - refer to the orientation of the M13 replication origin.	Stratagene; Mead, *et al.*, 1985
pBR329			Ap^r, Cm^r, Tet^r cloning vector comprising all of pBR322 plus a 1.2 kb fragment including the Cm^r gene.	Covarribius and Bolivar, 1982
p94.R1	pBR329	6.0 kb	Eco RI-Eco RI fragment of λ9405	J. A. Davies
p94.R2	pBR329	2.8 kb	Eco RI-Eco RI fragment of λ9405	J. A. Davies
p94.R3	pBR329	1.76 kb	Eco RI-Eco RI fragment of λ9405	J. A. Davies
p94.R4	pBR329	1.7 kb	Eco RI-Eco RI fragment of λ9405	J. A. Davies
p94.R5	pBR329	1.7 kb	Eco RI-Eco RI fragment of λ9405	J. A. Davies
p94.R7	pBR329	1.25 kb	Eco RI-Eco RI fragment of λ9405	J. A. Davies
p94.R8	pBR329	0.85 kb	Eco RI-Eco RI fragment of λ9405	J. A. Davies
KE2(1.8)	pBluescriptII KS+	1.8 kb	plasmid clone of λKE2 insert	Chapter 3
psB(1.8)RH	pBluescriptII KS+	0.25 kb	Eco RI-Hin dIII fragment of KE2(1.8)	Chapter 3
psB(1.8)HB	pBluescriptII KS+	0.8 kb	Hin dIII-Bam HI fragment of KE2(1.8)	Chapter 3
psB(1.8)BR	pBluescriptII KS+	0.77 kb	Bam HI-Eco RI fragment of KE2(1.8)	Chapter 3

pLethal	pBluescriptII KS-	1.04 kb	Larger Eco RI-Bam HI fragment of KE2(1.8)	Chapter 6
psBGA	pBluescriptII KS+	1.4 kb	Eco RI-Bam HI fragment from λ94C11	Chapter 3
psBGB	pBluescriptII KS+	0.37 kb	Hin dIII-Eco RI fragment from p94.R2	Chapter 3
psBGC	pBluescriptII KS+	0.8 kb	Bgl II-Bam HI fragment from p94.R1	Chapter 3
psBGD	pBluescriptII KS+	1.15 kb	Bam HI-Bgl II fragment from p94.R1	Chapter 3
psBGE	pBluescriptII KS+	1.76 kb	Eco RI-Eco RI fragment from p94.R3	Chapter 3
psBGE(i)	pBluescriptII KS+	0.97 kb	Eco RI-Pst I fragment from psBGE	Chapter 3
psBGE(ii)	pBluescriptII KS+	0.8 kb	Pst I-Eco RI fragment from psBGE	Chapter 3
psBGF	pBluescriptII KS+	1.7 kb	Tandem repeat of 0.85 kb Eco RI-Eco RI fragment from p94.R8	Chapter 3
psBGG	pBluescriptII KS+	1.2 kb	Bam HI-Hin dIII fragment from λ940	Chapter 3
psBGJ	pBluescriptII KS+	1.7 kb	Eco RI-Eco RI fragment from p94.R5	Chapter 4
psBGJRS	pBluescriptII KS+	0.5 kb	Eco RV-Sal I fragment of psBGJ	Chapter 4
psBGK	pBluescriptII SK-	1.7 kb	Eco RI-Eco RI fragment from λ 95204C9	Chapter 4
psBGKEH	pBluescriptII KS+	0.55 kb	Eco RI-Hin dIII fragment from psBGK	Chapter 4
psBGKHB	pBluescriptII KS+	0.38 kb	Hin dIII-Bam HI fragment from psBGK	Chapter 4
psBGKBE	pBluescriptII KS+	0.75 kb	Bam HI-Eco RI fragment from psBGK	Chapter 4
psBGM	pBluescriptII SK-	2.6 kb	Eco RI-Eco RI fragment from λ95204C9	Chapter 4
B10	pBluescriptII KS+	1.25 kb	cDNA fragment generated by PCR between PCR1 and NBT7 primers on NB12-24 hr embryonic cDNA library template	Marian Wilkin; Chapter 4
B10BEV	pBluescriptII KS-	0.31 kb	Bam HI-Eco RV fragment of B10	Chapter 4
B10EVB	pBluescriptII KS-	0.16 kb	Eco RV-Bam HI fragment of B10	Chapter 4

B10BB	pBluescriptII KS+	0.47 kb	Bam HI-Bam HI fragment of B10	Chapter 4
B10EVH	pBluescriptII KS-	0.35 kb	Eco RV-Hin cII fragment of B10	Chapter 4
B10HEI	pBluescriptII KS+	0.33 kb	Hin cII-Eco RI fragment of B10	Chapter 4
B10EIEI	pBluescriptII KS+	0.15 kb	Eco RI-Eco RI fragment of B10	Chapter 4
P1	pBluescriptII SK-	3.9 kb	Plasmid excised from λP1 cDNA clone	Chapter 3
P1PE	pBluescriptII KS+	1.2 kb	Eco RI-Pst I fragment of P1	Chapter 3
P1BP	pBluescriptII KS+	0.56 kb	Bam HI-Pst I fragment of P1	Chapter 3
P1KB	pBluescriptII KS+	0.24 kb	Kpn I-Bam HI fragment of P1	Chapter 3
P1EK	pBluescriptII KS+	1.28 kb	Eco RV-Kpn I fragment of P1	Chapter 3
P1XE	pBluescriptII KS+	0.6 kb	Eco RV-Xho I fragment of P1	Chapter 3
N52	pNB40	1.66 kb	cDNA clone isolated from NB12-24 hr library	Chapter 4
SIPC224	pNB40	0.97 kb	cDNA clone isolated from NB12-24 hr library by inverse PCR between P11 and P5	Chapter 4
SIPC726	pNB40	1.25 kb	cDNA clone isolated from NB12-24 hr library by inverse PCR between P11 and P5	Chapter 4
SIPC737	pNB40	2.15 kb	cDNA clone isolated from NB12-24 hr library by inverse PCR between P11 and P5	Chapter 4
SIPC8	pNB40	2.15 kb	cDNA clone isolated from NB12-24 hr library by inverse PCR between P11 and P5	Chapter 4
M23	pBluescriptII SK-	3.5 kb	Plasmid excised from λM23 cDNA clone	Chapter 4
M11B	pBluescriptII SK-	2.4 kb	Plasmid excised from λM11B cDNA clone	Chapter 4
M13	pBluescriptII SK-	1.5 kb	Plasmid excised from λM13 cDNA clone	Chapter 4
M12A	pBluescriptII SK-	2.45 kb	Plasmid excised from λM12A cDNA clone	Chapter 4

M14	pBluescriptII SK-	2.9 kb	Plasmid excised from λM14 cDNA clone	Chapter 4
S3A	pBluescriptII SK-	3.4 kb	Plasmid excised from λS3A cDNA clone	Chapter 4
M23XP	pBluescriptII KS +	0.45 kb	Xho I-Pst I fragment from M23	Chapter 4
M23PS	pBluescriptII KS +	1.1 kb	Pst I-Sma I fragment from M23	Chapter 4
M23SB	pBluescriptII KS +	0.66 kb	Sma I-Bam HI fragment from M23	Chapter 4
M23BK	pBluescriptII KS +	0. 2 kb	Bam HI-Kpn I fragment from M23	Chapter 4
pMGA	pBluescriptII KS+	0.8 kb	Hin dIII-Hin dIII fragment from λH962	Chapter 4
pMGB	pBluescriptII KS+	2.6 kb	Eco RI-Hin dIII fragment from λH962	Chapter 4
pMGC	pBluescriptII KS+	4.2 kb	Eco RI-Eco RI fragment from λH962	Chapter 4
pMGD	pBluescriptII KS+	1.5 kb	Eco RI-Hin dIII fragment from λH962	Chapter 4
pMGE	pBluescriptII KS+	1.6 kb	Sal I-Hin dIII fragment from λH962	Chapter 4
pMGF	pBluescriptII KS+	5.6 kb	Eco RI-Sal I fragment from λM954	Chapter 4
94C15R1	pBluescriptII KS+	6.2 kb	Eco RI-Eco RI fragment from λ94C15	Chapter 4

2.4 GEL ELECTROPHORESIS

Tris-borate-EDTA buffer (TBE):	0.1M tris, 0.89M H_3BO_3, 2mM Na_2EDTA, pH8.3
6x agarose gel loading buffer (AGLB):	0.25% (w/v) bromophenol blue, 0.25% (w/v) xylene cyanol FF, 40% (w/v) sucrose, (stored at 4°C)
2.5x sequencing gel loading buffer (SGLB):	95% (v/v) formamide, 20mM Na_2EDTA, 0.05% (w/v) bromophenol blue, 0.05% (w/v) xylene cyanol FF
40% (w/v) acrylamide-bis (29:1) solution:	38.7% (w/v) acrylamide, 1.3% (w/v) bisacrylamide, (stored in dark at 4°C)
Soak solution	10% (v/v) glacial HOAc, 12% (v/v) methanol

2.4.1 Agarose gels

Agarose gels (0.7% to 1.5% w/v) were prepared by adding TBE buffer to preweighed solid agarose, dissolving in a microwave oven, cooling until hand hot and supplementing with ethidium bromide (etBr) stock solution (10mg/ml) to a final concentration of 0.2µg/ml before casting. Gels were allowed to set at room temperature for a minimum of 1 to 3 hours depending on size and were submerged in TBE buffer before loading samples containing 1x AGLB. Electrophoresis was carried out at 1 to 10V/cm and was monitored by following the migration of the xylene cyanol and bromophenol blue dye markers. Gels were viewed on a Vilber Lourmat transilluminator emitting UV with a peak at 312nm. The above procedure was modified when DNA fragments were to be isolated for cloning or direct sequencing. In such experiments, gels were run in the absence of etBr and were later stained for 30 mins in a solution of 0.5µg/ml etBr in the dark, rinsed well in dH_2O and viewed on a prewashed long wave UV transilluminator (365 nm peak, 5.2mW.cm^{-2}, Ultraviolet Products Inc.) to allow gel fragments containing bands of interest to be excised with the minimum of DNA damage.

2.4.2 Sequencing gels

0.4mm thick, 6% acrylamide denaturing gels were used to resolve the products of sequencing reactions. Gels were made using 18ml 40% acrylamide-bisacrylamide stock, 12ml 10x TBE, 60g urea, and 36ml dH_2O per 100ml. The mixture was dissolved at 37°C, and filtered through a 0.45µm Minisart filter (Sartorius). Paired gel plates were cleaned thoroughly (3 washes with 2% (w/v) SDS, 3 washes with isopropyl alcohol (IPA), 3 washes with 70% (v/v) ethanol)

and assembled as a sandwich using 1 cm wide Whatman 3MM paper strips as spacers along the bottom and sides. The spacer edges were sealed with waterproof tape and the plates held together with 12 bulldog clips. The gel was poured using a 50ml syringe with the gel plates held at approximately 20° angles both along and across the plates. Once the gel was poured, a mylar sharkstooth comb (IBI) was inserted and clamped fast with bulldog clips. The gel sandwich was left at a 10° tilt (top to bottom), and left to set for 2 to 20 hours. After setting, clips and tape were removed, the sharkstooth comb inverted, and the gel pre-electrophoresed for 45 mins at 50W, using an LKB 2197 power supply. Wells were flushed out thoroughly with TBE buffer (to wash away urea leached from the gel), and heat denatured samples (2 mins at 80°C) containing 1x SGLB were loaded. Gels were electrophoresed at 40 to 60W (equivalent to 33 to 47 V/cm) for 2 to 5 hours. The gel plates were carefully separated, leaving the gel sticking to one or other surface. Gel and supporting plate were immersed in soak solution for 25 mins to remove the urea. The gel was then transferred onto a double thickness of 3MM paper, overlaid with Saran Wrap, dried at 80°C under vacuum on a Bio-Rad model 583 gel drier for 90 mins and exposed to X-Ray film.

2.5 *IN VITRO* MANIPULATION OF DNA

2.5.1 Elution of DNA from agarose gels

DNA was eluted from agarose gels by the method of Heery and coworkers (Heery, *et al.*, 1990). An agarose slice containing a DNA fragment of interest was isolated as described above (§2.4.1). The cap was cut off a 0.5ml microfuge tube, the bottom of the tube was pierced with a Microlance 25G needle (Becton-Dickson) and a plug of siliconised glass wool (about 100µl) was packed into the tube. The gel slice was placed on top of the glass wool plug, and the tube inserted into a 1.5ml microfuge tube which had also had its cap removed. The whole assembly was then centrifuged for 10 mins at 4000 rpm (896g) at room temperature, allowing the DNA solution to pass through to the lower tube while the agarose matrix was retained by the glass wool. The efficiency of DNA recovery was assessed by brief UV illumination of residue and filtrate. If a substantial fraction of the DNA remained associated with the agarose, a further 5 minute, 896g spin at room temperature was performed. The small tube was then discarded and the filtrate centrifuged for 5 mins at 13000rpm (9500g),

room temperature, to pellet any agarose fragments or glass wool fibres present. The DNA solution was extracted once against an equal volume of phenol:chloroform (1:1) and once against an equal volume of chloroform. The eluted DNA was then precipitated as described below (§2.5.5).

2.5.2 Restriction endonuclease digestion of DNA

DNA digestion with restriction endonucleases was carried out at DNA concentrations of 10-50µg/ml, and enzyme concentrations typically 5-10 units/µgDNA. 10x buffer solutions provided by the enzyme suppliers (BRL react™ , Pharmacia One-Phor-All™ and Promega 4 core™) were used. Reactions were incubated for 80 to 100 mins at 37°C, except SmaI digests which were incubated at 30°C, for a similar period.

2.5.3 Alkaline phosphatase treatment of DNA

For some experiments in which DNA fragments were ligated into vector molecules with self-complementary sticky ends, vector DNA was pretreated with calf intestinal alkaline phosphatase (CIP) to suppress the formation of self-ligated nonrecombinant molecules. CIP was added to restriction digests 20 mins before the end of the incubation time, at a concentration of 1U/100pmol 5' ends. Once the incubation was complete, the phosphatase was inactivated by heating to 80°C for 10 mins. Blunt ended vector molecules were not CIP treated.

2.5.4 Ligation of DNA fragments

10x ligation buffer:	300mM tris-HCl, 100mM $MgCl_2$, 100mM dithiothreitol (DTT), 10mM ATP, pH7.8

For intermolecular ligation reactions, around 50ng vector DNA were used and insert DNA was added with a vector:insert stoichiometry of between 1:1 and 1:8. The volume was adjusted to 8.5µl with dH_2O, and the solution heated to 45°C for 5 mins to melt cohesive termini, then cooled on ice. 1µl 10x ligase buffer and 1.5 Weiss units T4 DNA ligase were added and the reaction was incubated for 4 hr at 16°C (for cohesive ends) or 2 hr at room temperature (for blunt ends).

2.5.5 Precipitation of DNA

A variety of methods were used for DNA precipitation. That most frequently used is described here, while other methods are detailed in the protocols of which they form part. DNA was precipitated by adding 0.1 volume 3M NaOAc (pH5.2) and 2 volumes ethanol to a solution of DNA, mixing thoroughly, and incubating on ice for 15 to 30 mins. DNA was recovered by spinning for 30 mins at 11000g and 4°C. The supernatant was removed and the pellet washed with 1ml 70% (v/v) ethanol. After a further spin (5 mins, 11000g, 4°C), all supernatant was removed and the pellet air dried at room temperature or 37°C for at least 20 mins.

2.6 TRANSFORMATION OF *E. COLI* CELLS

Different methods were used for the preparation and transformation of *E. coli* cells, depending on the transformation efficiency required. The three protocols used most often throughout this work are presented in order of increasing transformation efficiency.

2.6.1 Preparation of competent *E. coli* cells

2.6.1.a Calcium chloride method

A 100ml culture of *E. coli* cells in 2xYT (§2.2.4) was grown overnight at 37°C with vigorous shaking. In the case of XL1-Blue cells, tetracycline was added to select for retention of the F' episome. 100ml prewarmed 2xYT without antibiotic were then inoculated with 4ml of overnight culture, and incubated at 37°C with vigorous shaking until the OD_{600} reached 0.3. The cells were then harvested by spinning at 5000rpm (3840g) in a precooled JA-14 rotor (Beckman) at 4°C for 5 mins, and the pellet resuspended in 50ml of 50mM $CaCl_2$ solution at 0°C. The suspension was incubated on ice for 30 mins, then the cells were pelleted as previously and resuspended in 10ml 50mM $CaCl_2$ solution at 0°C. Competent cells could be stored at this stage by adding 0.18 volumes sterile glycerol (0°C) and freezing at -70°C in 0.5 ml aliquots.

2.6.1.b Rubidium chloride method

TFB I:	30mM KOAc, 100mM RbCl, 10mM $CaCl_2$, 50mM $MnCl_2$, 15% (v/v) glycerol, pH5.8 (with HOAc)

TFB II:	10mM MOPS, 75mM $CaCl_2$, 10mM RbCl, 15% (v/v) glycerol, pH6.8 (with HCl)

A 100ml culture of *E. coli* cells was grown overnight in ψ broth (§2.2.4), with selective antibiotic if required. 100ml prewarmed ψ broth were inoculated with 4ml of overnight culture, and grown until the OD_{600} reached 0.46-0.6. The flask was then cooled on ice, and the cells harvested with a 10 minute spin at 5000rpm (3840g) in a JA-14 rotor (Beckman) precooled to 4°C. The pellet was resuspended in 20ml TFB I at 0°C, and kept on ice for 30 mins. The cell suspension was then transferred to a 40ml prechilled polypropylene tube and the cells were pelleted once more by spinning for 10 mins at 5000rpm (3020g) at 4°C in a precooled JA-20 rotor (Beckman). This pellet was then resuspended in 4ml TFBII at 0°C, and kept on ice for 15 mins. Cells could then be used directly for transformations, or stored. To store, cells were divided into 0.5ml aliquots in prechilled tubes and snap frozen by plunging into liquid N_2, then transferring to a -70°C freezer. Cells were found to retain competence for at least 8 months when prepared and stored in this way.

2.6.1.c Washing cells for electrotransformation

A culture of XL1-Blue cells was grown overnight in L. broth at 37°C, with vigorous shaking. 1ml of this culture was used to inoculate 400ml prewarmed L. broth. The 400ml culture was incubated at 37°C, with vigorous shaking, until a cell density of OD_{600} = 0.5 was reached, whereupon the flask was cooled on ice for 30 mins. The culture was divided into two precooled 250ml centrifuge bottles (Nalgene), and the cells were pelleted by spinning for 15 mins at 2460g, 4°C, in a precooled JA-14 rotor (Beckman). The cells were resuspended in 1l dH_2O at 0°C, harvested as above, resuspended in 500ml dH_2O at 0°C, harvested once more, then resuspended in 20ml 10% glycerol (v/v) at 0°C, and transferred to precooled 40ml polypropylene centrifuge tubes (Nalgene). The cells were pelleted again by spinning in a JA-20 rotor (Beckman) at 1940g, 4°C, for 10 mins, then resuspended in 2ml 10% glycerol (v/v) at 0°C. To store, cells were divided into 200µl aliquots in precooled 0.5ml microfuge tubes, snap frozen in a dry ice/ethanol bath and transferred to a -70°C freezer.

2.6.2 Transformation of competent *E. coli* cells

2.6.2.a Transformation of cells prepared according to the $CaCl_2$ (§2.6.1.a) and RbCl (§2.6.1.b) methods

A volume of 5µl or less of TE pH8.0 or 1x ligase buffer (§2.5.4) containing DNA for transformation was chilled on ice. 200µl competent cells were added and mixed in briefly. When frozen cells were used, they were thawed on ice then added immediately to the DNA solutions. In every case a control without vector DNA, but including all other components present in DNA solutions being transformed, was included to ensure that competent cells and ligation reaction components were free from resistant contaminating cells. Positive controls using 10ng circular pBluescriptII™ DNA were also included to enable estimation of transformation efficiency. DNA and competent cells were incubated together on ice for 30 mins. The cells were then heat shocked for 2 mins at 42°C, 750µl SOC (§2.2.4) medium were added, and the cultures were incubated for 30 minutes at 37°C with gentle agitation. Aliquots were then plated onto L. agar plates containing appropriate selective antibiotics and indicator reagents (§2.2.4), and incubated overnight at 37°C.

2.6.2.b Electrotransformation

Cells washed and frozen as described (§2.6.1.c) were thawed on ice. 40µl cell suspension were added to 1µl prechilled DNA solution in dH_2O and incubated for 5 mins on ice. (Control transformations without vector DNA and with intact plasmid DNA were included as described in §2.6.2.a.) The mixture was transferred to a chilled 0.1cm cuvette (Bio-Rad Laboratories) and subjected to a single 1.6kV, 25 µF pulse in a Gene Zapper (Bio-Rad Laboratories). 1ml SOC medium (§2.2.4) at 37°C was added to flush out the cells and the cell suspension was transferred to a sterile 20ml universal bottle and incubated, with shaking, for 1hr. Aliquots were then plated onto L. agar plates containing selective antibiotic and indicator reagents if required (§2.2.4). Plates were incubated overnight at 37°C.

2.7 HARVESTING DNA FROM *E.COLI* AND *DROSOPHILA*

2.7.1 Preparation of single stranded DNA from pBluescriptII™ clones

AAP Solution:	3.5MNH_4OAc, pH7.5, 20% (w/v) PEG-8000
TE pH 8.0:	10 mM tris-HCl, 1mM Na_2EDTA, pH 8.0

Single stranded DNA (ssDNA) was prepared according to the Stratagene pBluescriptII™ manual. 100ml superbroth (§2.2.4) containing ampicillin and tetracycline (§2.2.4) were inoculated with XL1-Blue cells carrying a clone of interest in pBluescriptII™, and incubated overnight at 37°C with vigorous shaking. A control experiment using pBluescriptII™ without insert was included to simplify the interpretation of the DNA species yielded, and to control against the presence of unrescuable sequences in the insert. Flasks containing 50ml prewarmed superbroth, supplemented with ampicillin, were inoculated with aliquots of overnight culture to a final cell density of OD_{600}=0.1, and incubated at 37°C with vigorous shaking. When the cell density reached OD_{600}=0.3 (equivalent to $2.5x10^8$ cfu/ml), helper phage (VCS-M13 or R408, Stratagene) were added at a multiplicity of infection of 20:1 (phage:bacteria). Incubation with vigorous shaking was continued for 8 hours, whereupon cultures were heated to 65°C for 15 minutes, transferred to polypropylene centrifuge tubes (Nalgene) and spun for 10 minutes at 11000g, 20°C in a JA20 rotor (Beckman). The supernatant was removed and spun again (10mins, 11000g, 20°C) to pellet any remaining cell debris. This supernatant was transferred to a fresh tube, and could be stored stably for several months at 4°C. ssDNA was purified by adding 1.2ml of supernatant to 300µl AAP solution in a microfuge tube, mixing and incubating at room temperature for 15 minutes. Phage pellets were recovered by spinning for 20 minutes at 11000g and room temperature. All supernatant was discarded, and the pellets resuspended in 20µl TE pH8.0. An equal volume of 1:1 (v/v) phenol:chloroform was added to the phage suspension and the mixture was vortexed for one full minute and spun for one minute. The upper, aqueous phase was then removed, leaving behind all of the interface. This procedure was repeated until no interface was detectable, then an extraction with an equal volume of chloroform was done to remove any traces of phenol. The ssDNA was precipitated by adding an equal volume of 7.5M NH_4OAc pH7.5 and two volumes of ice cold ethanol and incubating for 20 minutes on ice. ssDNA was recovered by spinning for 20 minutes at 11000g, 4°C. The supernatant was removed and the pellet rinsed with 1ml 80% (v/v) ethanol. After a final spin (2mins, 11000g) all liquid was removed, and the pellet left to air dry at room temperature. Once dry, the ssDNA was resuspended in 20µl TE pH8.0, and 5µl of each sample was run on a 1% agarose minigel. Yields varied between 0 and 2µg, depending upon the insert sequence.

2.7.2 Isolation of plasmid DNA from *E. coli*

Birnboim-Doly I (BDI):	50mM glucose, 25mM tris-HCl pH8.0, 10mM EDTA, 0.5% (w/v) lysozyme
Birnboim-Doly II (BDII):	0.2M NaOH, 1% SDS
Birnboim-Doly III (BDIII):	5M KOAc pH4.8
TE pH8.0:	10mM tris, 1mM Na_2EDTA pH8.0

2.7.2.a Small scale plasmid purification (Minipreps)

DNA was isolated using a modification of the procedure of Birnboim and Doly (Birnboim and Doly, 1979). 5 ml of L. broth containing the appropriate selective antibiotic were inoculated with a single *E. coli* colony and incubated overnight (at least 14 hr) at 37°C, with vigorous shaking. 1.5ml of the 5ml culture of stationary phase cells was harvested by centifugation at 9500g for 30 seconds and the pellet was resuspended in 100µl ice cold BDI solution. 200µl BDII were added with gentle mixing, followed by 150µl ice cold BDIII. Tubes were stored for 5 mins on ice, after vortexing briefly. Cell debris and most chromosomal DNA was then removed by centrifugation at 11000g, 4°C, for 5 mins. The supernatant was removed and residual protein contaminants extracted against 200µl phenol:chloroform (1:1). The aqueous layer was then extracted against 200µl chloroform to remove traces of phenol, and the nucleic acids precipitated (§2.5.5). The nucleic acid pellet was resuspended in 20µl TE pH8.0, containing 1mg/ml RNAase A.

2.7.2.b Large scale plasmid purification.

100ml superbroth containing the appropriate selective antibiotic were inoculated with a single *E. coli* colony and incubated overnight (at least 18hr) with vigorous shaking. Cells were harvested by centrifugation for 5 minutes at 5000rpm (3840g), 4°C, in a JA-14 rotor (Beckman). The pellet was resuspended in 5ml BDI, and incubated for 5 mins at room temperature. 10ml BDII were added and the bottle left on ice for 10mins. 7.5ml ice cold BDIII were added and, after a 10 minute incubation on ice to allow the precipitation of proteins and chromosomal DNA, this debris was pelleted by centrifugation for 20 minutes at 9000rpm (12400g) and 4°C in a JA-14 rotor. The supernatant was transferred to a 50ml Falcon tube and the nucleic acids precipitated by adding 0.6 volumes IPA and spinning for 30 minutes at 4000rpm (2420g), 20°C, in a Jouan benchtop centrifuge. The nucleic acid pellet was washed with 10ml 70% v/v ethanol at room temperature, air dried briefly, and resuspended in 5ml TE

pH8.0. Exactly 4.5ml of this solution were added to exactly 4.50g CsCl in a 20ml universal bottle. 200µl etBr solution at 10mg/ml were added and the mixture transferred to a Beckman 3.5in polyallomer ultracentrifuge tube using a pasteur pipette. The tube was sealed and spun at 55000rpm (267000g) in a vTi65 rotor (Beckman) for 6-8 hours at 20°C. On viewing over a long wave (365 nm peak) UV source, one or two bands were visible, a major band of supercoiled DNA always being prominent. Up to 700µl of solution containing this band were removed with a microlance 21G needle (Becton-Dickson) and sterile syringe, after an air bleed needle had been passed through the shoulder of the tube. The etBr was removed by a series of up to 10 extractions against equal volumes of CsCl saturated IPA. Dialysis over 24 hr against 3 changes of TE pH8.0 was performed to remove CsCl. The dialysis tubing used was 9/32" wide Visking membrane (Medicell International). Quality and yield of DNA were assessed by running small aliquots of uncut and restriction enzyme cut DNA on 0.8% agarose gels alongside known quantities of cut bacteriophage λ DNA. Yields from 100ml XL1-Blue cultures containing pBluescriptII™ plasmid clones were typically 500-1000µg.

2.7.3 Preparation of genomic DNA from *Drosophila*

Homogenisation buffer:	10mM tris-HCl pH8.0, 60mM NaCl, 10mM Na_2EDTA pH8.0, 150µM spermine, 150µM spermidine, 0.5% (v/v) Triton X-100
Lysis buffer:	Homogenisation buffer, 0.4% (w/v) SDS
TE pH8.0:	10mM tris-HCl, 1mM Na_2EDTA, pH8.0

Flies from which DNA was to be prepared were starved for 4 hours at room temperature in a population cage containing a moistened tissue as a humidifier, but no food. This was done to minimise the quantity of DNA harvested from ingested yeast. Around 10 flies were homogenised in 200µl homogenisation buffer in a 1.5ml Eppendorf SafeLock™ tube, using a pestle homogeniser attached to a Bosch household drill. 200µl lysis buffer were added, followed by proteinase K to a final concentration of 100µg/ml (2µl 20mg/ml stock solution). The reactions were mixed gently, incubated at 37°C for 1.5 hr, then extracted with equal volumes of phenol, phenol:chloroform (1:1) and finally chloroform. Nucleic acids were precipitated (§2.5.5) and the air dried pellets resuspended in 50µl TE pH8.0, 0.4% (w/v) RNAse A. DNA quality and yield were assessed by running aliquots on a 0.8% agarose gel.

2.8 SYNTHESIS OF LABELLED DNA

2.8.1 Preparation of ^{32}P labelled random primed DNA probes

Solution A:	1.25M tris-HCl pH8.0, 0.125M $MgCl_2$, 1.8% (v/v) β-mercaptoethanol, 0.5mM dATP, 0.5mM dGTP, 0.5mM dTTP
Solution B:	2M HEPES pH 6.6
Solution C:	Random hexanucleotide primers at 90 OD units/ml in TE pH7.5
Oligonucleotide Labelling Buffer (OLB):	Prepared by mixing solutions A, B and C, in the ratio 2:5:3
SE solution:	0.5% SDS, 10mM Na_2EDTA pH8.0

The method used was a slight modification of the procedure of Feinberg and Vogelstein (Feinberg and Vogelstein, 1983; Feinberg and Vogelstein, 1984). 10 to 100ng purified DNA in 30µl dH_2O or TE were boiled vigorously to denature DNA strands, and immediately snap cooled on ice. Water was added such that the final volume of the reaction would be 50µl, followed by 10µl OLB, 15-40 µCi $\alpha^{32}P$ dCTP and 2U Klenow polymerase. The reaction was incubated at room temperature for 10-16 hours. Spermine precipitation was used to enable removal of unincorporated nucleotides: 2.5µl 0.2M spermine were added, the reaction was left on ice for 30-40 mins and then centrifuged for 15 mins at 11000g and 4°C. The supernatant was removed and transferred to a fresh tube. Percentage incorporation of labelled nucleotide was estimated by monitoring the supernatant and pellet fractions with a hand-held Geiger-Muller counter. The pellet was resuspended in 500µl SE solution and stored at -20°C until required.

2.9 MEMBRANE HYBRIDISATIONS

2.9.1 Southern transfer of DNA from agarose gels to Hybond™ C-Extra membranes

Bidirectional Southern transfers were carried out according to the method described by Sambrook *et al* (1989) after Smith and Summers (Smith and Summers, 1980; Southern, 1975).

Denaturing solution:	0.5M NaOH, 1.5M NaCl
Neutralising solution:	1M tris-HCl pH7.4, 1.5M NaCl
20x SSC:	3M NaCl, 0.3M Na_3citrate, pH7.0

An agarose gel was prepared and run as described in §2.4.1. After photography, unused areas of gel, including the region above the gel slots, were trimmed away, and the top left corner was removed as an orientation marker. The gel was then immersed in several volumes of 0.25M HCl for 20 mins to partially depurinate DNA. The gel was rinsed in dH_2O and transferred into denaturing solution for 30 mins. This alkaline solution both denatures DNA strands and cleaves the phosphate backbone at apurinic sites generated by HCl treatment. After a further dH_2O rinse, the gel was transferred into several volumes of neutralising solution for 50 mins, then to 20x SSC for 10 mins. Six sheets of Whatman 3MM paper and two of Hybond™ C-Extra of the same size as the gel were cut. The nitrocellulose sheets were wetted in dH_2O. One sheet was laid on top of a triple thickness of 3MM paper sheets, and the gel was placed on top of the membrane, with care being taken to avoid bubbles. A strip of parafilm was inserted between each edge of the gel and the nitrocellulose, overlapping the gel by about 1mm. This served to isolate the gel from the blotting paper and paper towels such that all fluid transfer occurred through the membrane. A second nitrocellulose filter was laid on top of the gel, and another parafilm border set in place. The upper filter was then overlaid with three sheets of 3MM paper, and the whole gel sandwich placed between two 4cm thicknesses of paper towels. A large glass plate and 400g weight were placed on top to aid capillary action, and the gel was blotted for at least 12 hrs. After this time, the gel sandwich was dismantled, and the filters labelled with an indelible pen. Filters were rinsed for 5 min in 2x SSC, then baked at 80°C for 2 hrs.

2.9.2 Transfer of bacterial colony DNA to Hybond™ C-Extra membranes

Denaturing solution:	0.5M NaOH, 1.5M NaCl
Neutralising solution:	1M tris-HCl pH7.4, 1.5M NaCl
Solution S:	1.8M NaCl, 0.03M Na_3citrate, pH7.0

A slight adaptation of the method of Grunstein and Hogness (Grunstein and Hogness, 1975) was used to fix bacterial colony DNA to Hybond™ C-Extra membrane circles. XL1-Blue cells containing recombinant plasmids of interest were grown overnight at 37°C on antibiotic-supplemented L. agar plates, either

after direct plating of cells at low density (<300 colonies per plate) or after patching out colonies individually into a regular array. A dry, numbered, circular Hybond™ C-Extra filter was placed on the surface of each plate to be screened. Orientation marks were made by punching needle holes at asymmetric points through the filter and the agar. The positions of these holes were also marked in pen on the underside of each petri dish. When completely wetted, each filter was lifted gently from its plate with blunt forceps and placed colony side up on a double thickness of Whatman 3MM paper saturated with denaturing solution. After five minutes, filters were removed from the denaturing solution, blotted briefly on a dry paper towel, then transferred onto a double thickness of 3MM paper soaked in neutralising solution. After five minutes, filters were again blotted on dry towel to remove excess solution, then were transferred onto 3MM paper soaked in solution S. After a further five minute incubation, filters were blotted on a paper towel, air dried for 1 hour on the bench, then baked at 80°C for 2 hours to bind the colony DNA firmly to the membrane. In order to ensure that live cells corresponding to positive colonies could be recovered, plates were incubated at 37°C for 1-6 hours (depending on the colony size remaining) after the filter lifts.

2.9.3 Hybridisation to membrane bound DNA

20x SSPE:	per litre: 175.3g NaCl, 27.6g $NaH_2PO4.H_2O$, 7.4g EDTA pH7.4 with NaOH
50x Denhardt's solution:	1% (w/v) BSA, 1% (w/v) Ficoll, 1% (w/v) PVP (M_r 360000)
Prehybridisation solution:	5x SSPE, 5x Denhardt's solution, 100µg/ml heat denatured, sonicated salmon sperm DNA, 0.5% (w/v) SDS
2x SSC:	300mM NaCl, 30mM Na_3citrate, pH 7.0
Wash A:	2x SSC, 0.1% (w/v) SDS
Wash B:	0.5x SSC, 0.1% (w/v) SDS

Filters from Southern blots (§2.9.1) or colony lifts (§2.9.2) were washed briefly in 2x SSPE, and sealed into plastic bags containing around 300µl prehybridisation solution for every cm^2 of filter area. Filters to be hybridised with the same probe were sealed into the same bag, up to a maximum of six small filters. Prehybridisation was done at 65°C for 1-2 hours, with shaking. An aliquot of ^{32}P labelled, random primed probe (§2.8.1) was heat denatured in a heating block at 100°C, and snap cooled on ice. The denatured probe was added to the hybridisation bag, which was then resealed and shaken briefly to mix. Filters were hybridised at 65°C with shaking for at least 12 hours, then washed

for 15 minutes at room temperature with wash A. Three 45 minute washes in wash A at 65°C and one 30 minute wash in wash B at 65°C were then done to remove nonspecifically bound probe. After rinsing in 2x SSC at room temperature, filters were blotted briefly on Whatman 3MM paper, mounted on a fresh sheet of 3MM paper and fixed in place with self-adhesive tags. ^{32}P labelled dye was spotted in an asymmetrical pattern onto the 3MM paper to allow orientation of the autoradiograph. The mounted filters were then wrapped in Saran Wrap and exposed at -70°C, with intensifying screens, to autoradiography film. Autoradiographs were developed as described in §2.14.2.

2.10 OLIGODEOXYRIBONUCLEOTIDES

Oligodeoxyribonucleotides, hereafter referred to as primers, were used extensively for both DNA sequencing and polymerase chain reactions. T3 and T7 promoter primers were obtained from Promega. All others were synthesised with a PCR mate (Applied Biosystems) using standard phosphoramidite chemistry (Caruthers, 1985; Gait, 1990).

2.10.1 Design of primers

Potential primer sequences were tested for their ability to form primer dimers (due to 3' end complementarity with themselves or other primers) using the BESTFIT (Devereux, *et al.*, 1984) program. Possible secondary structure formation was investigated using the FOLD (Zuker and Stiegler, 1981) program. BESTFIT and FOLD are both constituents of the Wisconsin GCG sequence analysis package (Devereux, *et al.*, 1984). Melting temperature (T_m) was estimated using the OLIGO 3.0 program (Rychlik and Rhoads, 1989) run on an Epson Equity II PC. This program uses very accurate measurements of nearest neighbour ΔG values (Breslauer, *et al.*, 1986; Freier, *et al.*, 1986) to derive an estimate of the melting temperature (T_m) of a primer of given sequence. Since T_ms are calculated for conditions of 50 mM KCl, a correction term had to be included when buffer conditions deviated from this, according to the equation:

$$T_m = \frac{\Delta H}{\Delta S + R.\ln(c/4)} - 273.15 + 16.6\log_{10}[\text{salt}]$$

where ΔH and ΔS are enthalpy and entropy for double helix formation respectively, R is the molar gas constant (1.987 cal/°C.mol), and c is the primer concentration (Schildkraut and Lifson, 1965).

2.10.2 Dissociation and deprotection of primers

On removal of the synthesis column from the PCR mate™, the primer is covalently bound via a succinate linkage to a solid support of controlled pore glass (CPG). Furthermore, the phosphoramidite chemistry used in the synthesis leaves backbone phosphates protected by cyanoethyl groups, which must be removed prior to use. Primers were detached from the CPG support and deprotected as follows. The synthesis column was broken open and the beads transferred to a 1.5ml microfuge tube. 1ml 40% (w/v) NH_4OH solution was added. After vortexing, the mixture was incubated on the bench for 1-2hr. During this incubation, the succinate linkage attaching the oligonucleotide to the CPG support is cleaved. The resulting solution of oligonucleotide in 40% NH_4OH was transferred to a 2ml screw capped tube (Nunc), leaving behind all of the CPG support. A further 1ml 40% NH_4OH solution was added and the tube was sealed and incubated at 55°C overnight, allowing the complete removal of the cyanoethyl protecting groups by a β-elimination reaction. An aliquot of the primer solution was precipitated by adding 0.1 volume 5M NH_4OAc pH7.0 and 2 volumes ethanol, incubating on ice for 30 mins, then spinning at 11000g, 4°C for 30 mins, washing the pellet with 1ml 70% (v/v) ethanol and air drying at room temperature. The DNA was resuspended in 100µl TE pH7.4. 2.0µl of the primer solution were added to 998µl TE pH7.4 and the OD_{260} of the resulting solution measured using a Beckman DU-50 spectrophotometer. The concentration was calculated by applying the Beer-Lambert law, taking ε_{260} as 12010 for G, 15200 for A, 8400 for T and 7050 for C, and calculating the ε_{260} for a primer by summing the contributions of each base (Wallace and Miyada, 1987). Typical total yields for a 20mer were 600 to 1000µg.

2.10.3 Catalogue of primer sequences

Primer Name	Sequence (5'-3')	Predicted T_m (°C)[1]

[1] Values given here refer to conditions of 0.05M KCl.

P1	CCAAGCTCAG ATAAATCAC	54.5
P2	CTGTTTTAAC GTGACTGACC	57.4
P3	GACCAATTGC AGCCAAATGC	68.3
P4	GTAAATTAGC AGCGGATTCG	62.7
P5	TCGAGTCCCG ATTAACCGAT	66.5
P6	GATTGAGCTA CCGAATCCTG	61.8
P7	CAATTTGCAC AACTTGCCCA	67.3
P8	CCGTTTGGAT TTGTTGGCCT	69.2
P10	GCTGACGACA AACCATG	56.5
P11	CTTCGGGTCC GTTCCTC	62.5
P12	CTGGCCAAGA ATTGAGG	58.2
PA1	CGGGTCTTCG TAGGGTTGTC	66.4
PD1	AAGGGATGGT GCCACTC	59.7
PD2	ATCATTGGTA CCAAGCTAAC A	59.6
PD4	GCTCATCCTC TTAAAAGG	53.8
PE1	TCGGGGAAAT GCAATTGGCG	75.2
PE2	CACATCTGAA AGACGACGTC	60.3
PE3	TACTTACTTGTAGTTTGTGG	49.7
PE4	CTTGTATCGT AAATTAGCAG	52.9
PF1	TTACTTCAAT ATTGGGAGG	54.6
PF2	GCTGGGGAAG GTCAGTTGTT T	67.6
PF3	CAGTTCCAAT GCGACATTAC GC	70.0
PF4	GGAAATTCAA TTACTTCAAT ATTGGGAGG	72.0
PF5	TTTGTTCCGA AGCCGAGAAA	67.8
PJ1	GCACATGACG ATCTCAGGAC	62.6
PJ2	ATCTGAATCA CAGCTAATGG CT	63.0
PK1	CATAGAAACG CAAATCCTTA GC	63.8
PK2	AAGTAGGACA GATGAATAGG CAGA	64.1
PK3	GCTAATATAC AGGGTGGTTA TTATAT	60.6
PK4	CCAGAATGAC CAAATGTACA TTT	63.3
PK5	GTTAATAATG TGAGAAACGT TAATAAT	60.9
PK6	GCCCAGGTTT CTTGTCAAT	61.7
PK7	CGTGATGTTG TACAGGACTT	57.3
PK8	TATAACTGTA TCTATGTTTT CACC	55.9
PM1	TTAATGAAGC GGCTGCTG	62.7
PUP1	GGCGGAGAAA CTGCGGATT	69.4
PCR1	CCAGTACGTT GGCAATCCGA TCGAT	76.8
PCR2	GCCTATGTCT AAGTCCATGA TGAGC	68.3
PUTR1	TTATGATCAA GCGGTGGCCT GC	73.8
PUTR2	GCGTTGGCAA AGTGAACGTG CG	77.2
PORF1	TGATTCGACA ACGACTATCG GT	66.7
PORF2	CGATACGCAA CGCGAACCAA TC	74.2
P1P1	AGTGTCTAAATGCAGAAACAA	58.2
P1XEP1	ATTCAGCAAC AGCCTTG	55.9
P4AP1	GCAGGAAGCC AAACAGG	61.2
P8P1	GGAAATCAAG TGCTAGCAC	57.3

P8P2	ACTGGAGAACAACTGTTTCC	58.2
NB1	GGAATTCCGG TGACACTATA GAATACAAGC TTGC	65.9
NBT7	AATACGACTC ACTATAGGGA GACCG	67.2
T3	ATTAACCCTC ACTAAAG	45.0
T7	TAATACGACT CACTATAGGG	51.7
GT10FOR	GAGCAAGTTC AGCCTGGTTA AGTC	68.5
GT10REV	GGCTTATGAG TATTTCTTCC AGGGTA	68.6
M23PKP0	AGGTTGCGGG TCAGCAG	64.2
M23PKP1	CCTTCAGGAC CATTGGC	60.5
M23PKP2	GATTCGGAAG CACTGCCG	66.7
M23PSP1	ATCAACAACA CCGCCGT	61.5
M23PSP2	CTCGATTCGC CACGGAG	64.8
M23PSP3	GAGTGAAGGC CGCAATC	60.9
M23PSP4	TGTCTGGCTA TCAAGAG	49.2
M23PSP5	GGTTAAACCG GATGCTG	58.4
M23PSP6	CTCGTCGTTG CATCGGC	65.7
M23PSP7	GACCCTTTTG ATACTAGT	47.0
M23PSP8	CAGCGTACGG TATTGCA	58.2
M23SBP1	GCGCTGCGCA GGGATCT	69.8
M23SBP2	CAAGTAAGCA ATATAGG	44.1
M23SBP3	CGCGCTGGTC ACGCAAC	69.6
M23SBP4	TAAGCCGGAC ACGCTCA	63.2
M21P1	GGGTGGCAAG TGCTCAAC	63.6
M21P3	AGTGTATAAG TTGTGGT	41.3

2.11 DNA SEQUENCING AND DATA ANALYSIS

2.11.1 Sequencing strategy

A directed strategy involving subcloning of restriction fragments and sequencing from vector based and insert-dedicated primers was adopted. Dedicated primers were used to sequence across subclone boundaries in parent clones, in order to verify that subclone sequences were contiguous and that small restriction fragments had not been lost. KE2(1.8), B10, and M23 cDNA sequences were derived on both DNA strands. Sequences of other cDNA clones and of genomic DNA fragments were derived on one or both strands.

2.11.2 DNA sequencing reactions

DNA sequencing reactions were carried out using the Sequenase™ version 2.0 enzyme, and according to the methods published in 'Step by Step Protocols for Sequencing with Sequenase™ Version 2.0' (United States Biochemicals).

5x Sequenase™ buffer:	200mM tris-HCl pH 7.5, 100mM $MgCl_2$, 250mM NaCl
5x dGTP labelling mix:	7.5µM dGTP, 7.5µM dCTP, 7.5µM dTTP
ddATP termination mix (for dGTP):	80µM dGTP, 80µM dATP, 80µM dCTP, 80µM dTTP, 8µM ddATP, 50mM NaCl
ddCTP termination mix (for dGTP):	80µM dGTP, 80µM dATP, 80µM dCTP, 80µM dTTP, 8µM ddCTP, 50mM NaCl
ddGTP termination mix (for dGTP):	80µM dGTP, 80µM dATP, 80µM dCTP, 80µM dTTP, 8µM ddGTP, 50mM NaCl
ddTTP termination mix (for dGTP):	80µM dGTP, 80µM dATP, 80µM dCTP, 80µM dTTP, 8µM ddTTP, 50mM NaCl
5x dITP labelling mix:	15µM dITP, 7.5µM dCTP, 7.5µM dTTP
ddATP termination mix (for dITP):	80µM dITP, 80µM dATP, 80µM dCTP, 80µM dTTP, 8µM ddATP, 50mM NaCl
ddCTP termination mix (for dITP):	80µM dITP, 80µM dATP, 80µM dCTP, 80µM dTTP, 8µM ddCTP, 50mM NaCl
ddGTP termination mix (for dITP):	160µM dITP, 80µM dATP, 80µM dCTP, 80µM dTTP, 8µM ddGTP, 50mM NaCl
ddTTP termination mix (for dITP):	80µM dITP, 80µM dATP, 80µM dCTP, 80µM dTTP, 8µM ddTTP, 50mM NaCl
2.5x SGLB:	95% (v/v) formamide, 20mM Na_2EDTA, 0.05% (w/v) bromophenol blue, 0.05% (w/v) xylene cyanol FF
Enzyme dilution buffer:	10mM tris-HCl pH7.5, 5mM DTT, 0.5mg/ml BSA

2.11.2.a Annealing of primer to single stranded templates

0.5-1.0µg of single stranded DNA in dH_2O or TE was mixed with 5ng primer. The volume was adjusted to 7µl with dH_2O, then 2µl Sequenase™ buffer and 1µl dimethylsulphoxide (DMSO) were added. The reaction was then heated to 80°C and allowed to cool to below 30°C over a period of approximately 30 minutes, then stored on ice until required.

2.11.2.b Denaturation and annealing of double stranded templates

This was done by the method of Zhang and coworkers (Zhang, *et al.*, 1988). 2-4µg double stranded DNA in 20µl 0.2M NaOH were incubated for 30 minutes at 37°C, neutralised with 2µl 5M NH_4OAc, pH 4.6 and precipitated by immediately adding 50µl ice-cold ethanol and leaving on ice for 15 minutes.

Denatured DNA was recovered by spinning at 11000g, 4°C, for 20 minutes, washing with 70% (v/v) ethanol, and air drying. To the dry DNA, 1µl (5ng) primer, 6µl dH_2O, 1µl DMSO and 2µl Sequenase™ reaction buffer were added, and the mixture incubated for 30 minutes at 37°C, then stored on ice until required.

2.11.2.c Labelling and termination reactions

To the annealed primer-template solution were added 1µl 0.1M DTT, 2.0µl 1x dGTP labelling mix, 0.2µl (2µCi) $\alpha^{35}S$ dATP (≈ 1100 Ci/mmol) and 2.0µl Sequenase™ version 2.0, diluted 9 fold in enzyme dilution buffer. This was carefully mixed, avoiding bubbles, and incubated at room temperature for 2 minutes. Tubes containing 2.5µl of each termination mix for dGTP were set up and prewarmed to 42°C. After the two minute incubation, 3.5µl of labelling reaction were transferred to each tube and the termination reactions were incubated at 42°C for 5 minutes. 4µl 2.5x SGLB were then added to stop the reactions, and the samples were analysed on 0.4mm denaturing polyacrylamide gels as described (§2.4.2).

2.11.3 Analysis of DNA sequence data

The Wisconsin GCG version 7.0 package (Devereux, *et al.*, 1984), run on a DEC Microvax 3600, was used for the analysis of DNA sequence data. Macvector (IBI) and Macpattern (Fuchs, 1991) were also used for the analysis of peptide sequences, as was the *Signify* program, which was designed to search for putative signal sequences within eukaryotic proteins (see below and Chapter 5). Other Macintosh programs used include the *Plota_KKD*, *Plota_TMH*, and *Plota_HYD* programs from the *MacProt* suite written by Annette Luettke and Peter Markiewicz. These were obtained from the EMBL fileserver (Fuchs, *et al.*, 1990). The *Top_Pred 3.2* program of Laszlo Sipos and Gunnar von Heijne (Sipos and von Heijne, 1993) was also extensively used. Macintosh programs were run on Macintosh LC, LCII and IIci personal computers.

Of the many sequence data analysis programs used, I will discuss four in detail. *Testcode* and *Codonpreference* analyses will later be used to anticipate which of several potential coding segments of cDNAs are indeed translated. Such predictions are both important and somewhat more contentious than most sequence data analyses, thus requiring both justification of their use and

assessment of their heuristic powers. The use of these programs is discussed in the following sections.

The *Fasta* program (Pearson and Lipman, 1988) was used to search sequence databases for potential homologues, while *Rdf2* (Pearson and Lipman, 1988) was used to calculate *z* scores to test the significance of the search results (see §2.11.3.b, below). Database searching is a discipline which tends often to be subject to the whims of the investigator. I have attempted to be systematic and consistent with database searches, as described below.

The *Signify* program, which predicts the presence of signal peptides in eukaryotic proteins was written specifically to address the question of whether any Shak-B proteins might be secreted or associated with the plasma membrane. This program is described in chapter 5.

2.11.3 Identification of coding regions in DNA sequences

2.11.3.a.i *Testcode*

Testcode (Fickett, 1982) is an empirical measure of the likelihood that a given DNA sequence encodes a peptide product. The algorithm surveys a window of DNA sequence (a 200 base window is the smallest found to give reliable results, and is used here), and evaluates eight parameters from the sequence. Four of these are the *position parameters*. The 'A' position parameter is derived from calculating how many times A appears in positions 1, 4, 7, 10 ..., how many times A appears in positions 2, 5, 8, 11..., and how many times A appears in positions 3, 6, 9, 12 ..., then dividing the largest of these three numbers by (the smallest + 1). Position parameters of the other bases are derived in an identical fashion. It is found that coding sequences, as a consequence of having non-random use of codons, have higher position parameters than non-coding sequences. The remaining four parameters used by *Testcode* are simply the number of times that each base appears within the sequence. While the relationship between base composition and probability of coding is rather subtle, it was found that certain percentage compositions were found more or less often in coding versus non-coding DNA. For example, a sequence which has a 17-19% T content, was found by Fickett's survey to have an 81% probability of being coding. By observing the distributions of each of the eight parameters in coding and noncoding DNA, Fickett demonstrated the relative

predictive value of each parameter and weighted them accordingly to generate the *Testcode* statistic. It is important to note that *Testcode* does not require codon usage data, but it does depend heavily on the base position parameters which will be highest in coding regions with high codon bias. Thus *Testcode* and *Codonpreference* measures are distinct but non-independent.

2.11.3.a.ii *Codonpreference*

The genetic code is degenerate, but synonymous codons are not used with equal frequency. Under the *genome hypothesis* formulated by Grantham and coworkers (Grantham, *et al.*, 1981; Grantham, *et al.*, 1980), bias in codon usage is species-specific, at least in lower organisms (including *Drosophila*) where the selective pressure underlying the evolution of nonrandom codon usage is thought to be the different abundances of synonymous tRNAs (Ikemura, 1981). However, other workers have demonstrated that enormous intraspecific variation in codon usage exists, and have observed a positive correlation between degree of bias and abundance of protein product for genes within a species (Gouy and Gautier, 1982; Sharp, *et al.*, 1988).

In trying to assess the probability that an ORF is genuinely coding, it is valid to ask to what extent the observed codon usage pattern mimics the expected codon bias of that organism. Under the original genome hypothesis, all that is required is to compile a vast collection of (in our case) *Drosophila melanogaster* coding sequences, recording the frequency with which each synonymous codon is used, and to compare these frequencies with those found in our ORF. A compilation of synonymous codon frequencies found in *Drosophila* sequences has been assembled by Michael Ashburner and is noble in its magnitude. However, it is worth questioning the wisdom of using codon frequency data amassed indiscriminately from *Drosophila* sequences, given the existence of wide intraspecific codon bias variations.

Consider the case of two novel, *bona fide Drosophila* coding sequences, one exhibiting high codon bias, the other showing low bias. Comparing codon usage of the high bias gene to that of a database of sequenced *Drosophila* genes might give only a moderate codon preference score, as low bias genes in the database could mask the codon usage patterns of the high-bias sequences. Conversely, a low-bias *Drosophila* gene may not be detected as coding because of the influence of the high-bias sequences in the database. The extent to which

these effects will be seen depends on the relative numbers of high and low bias genes used to compile the codon usage data.

An alternative approach is to subdivide known *Drosophila* coding sequences according to their degree of codon bias, and to compare the codon usage patterns of our query sequences with the patterns observed in high, medium or low bias *Drosophila* sequence databases. Such an approach would be expected to successfully identify a highly biased *Drosophila* test sequence. Whether the same is true of low bias sequences is less clear. If translational selection is the only evolutionary force underlying *Drosophila* codon bias, then low abundance genes might be expected to tend towards equal use of synonymous codons, as would appear in non-coding sequences. Thus a low-bias data set might not enable us to distinguish between low-bias coding and random non-coding sequences. It is possible, however, that in the absence of strong translational selection, mutational bias could maintain a lower level of codon usage bias. Indeed, the pattern of codon usage in weakly expressed *B. subtilis* genes is highly indicative of mutational bias (Shields and Sharp, 1987), though there is not currently any evidence for mutational bias in *Drosophila* coding sequences. In this work, I have therefore tested all sequences only with codon frequency data from a high-bias *Drosophila* sequence compilation, using data presented by Shields *et al*, (Shields, *et al.*, 1988). This is an effective way to assess whether a test sequence shows codon usage characteristic of highly biased *Drosophila* genes and avoids the loss of resolution which would result from the common practice of using of codon data from genes with heterogeneous degrees of bias.

The codon preference plotting program of Gribskov and coworkers (Gribskov, *et al.*, 1984) is used in this work. The plot is constructed by calculating a codon preference statistic at each position for each of three reading frames. The statistic is calculated across a window of specified size (25 codons in this work), and the window moves in single codon increments, thus generating three plots, one for each frame. For each codon in the test sequence, the relative likelihood (from codon usage data) that the codon comes from a genuine coding sequence is compared to the likelihood that it would arise randomly in a scrambled sequence of the same base composition. This has the effect of making the statistic robust to fluctuations in test sequence base composition. Unlike some of its predecessors (Staden and McLachlan, 1982) the method of Gribskov *et al* assesses only the relative abundance of synonymous codons, and ignores the

relative abundance of different amino acid residues, thus allowing fair codon usage assessment of sequences with unusual amino acid compositions.

2.11.3.b Database searching

All database homology searches were done using the *Fasta* and *Tfasta* programs (Pearson and Lipman, 1988), as implemented in the Wisconsin GCG package release 7.0. All sequences obtained from *shaking-B* and its environs were used as query sequences for DNA level and translated DNA level searches against the GenBank, EMBL, SwissProt and PIR databases. Most initial database searching was done against those releases available locally in Glasgow. In order to access the most recent versions of the sequence databases, query sequences were submitted to the EMBL Fasta server and searches were run against databases updated daily (Fuchs, *et al.*, 1990). DNA searches were all performed using a word size (k-tuple) of 6, in other words, the first step of the comparison involved finding matches between each database sequence and each six base window of the query sequence. The search is subsequently refined to generate an optimised similarity score, and part of this refinement involves comparison of the query sequence and identified homologues using windows smaller than the prescribed k-tuple. However, since the *Fasta* algorithm depends initially on finding blocks of sequence identity, the bigger the k-tuple the greater the number of potential homologues which will be thrown out before the search refinements begin. The smaller the k-tuple, the higher the number of comparisons done and, effectively, the more sequences scored for homology. However, this greater sensitivity incurs great penalties in the processing time required to do the search, and, given a limited amount of cpu time, 6 was the minimum k-tuple used for DNA level searching.

In comparisons designed to find homologous structural genes, searches at the protein level are much more sensitive than those at the DNA level. For those sequences proposed to encode proteins (see above), the search strategy was to translate the sequence and search initially with an amino acid k-tuple of 2, only moving on to searches with a k-tuple of 1 for those query sequences which yielded no homologues in less sensitive searches and yet were confidently identified as coding.

The question of statistical validity of database search results is a vexed one, which has been tackled lucidly by only a few authors (Karlin, *et al.*, 1983;

Lipman and Pearson, 1985; Lipman, *et al.*, 1984; Sankhoff and Cedergren, 1973; Steele, 1982). Sadly, similarity scores generated by *Fasta* and *Tfasta* do not follow a normal distribution from which probabilities can be calculated. Once a database-resident sequence has been identified as a homologue, it must be either a genuine homologue or a false positive due to insufficient selectivity in the search. The primary cause of false positives in database searching is locally high concentrations of particular amino acids or nucleotides (Lipman, *et al.*, 1984) and the conventional way to circumvent this problem is by comparing scores of *query sequence vs potential homologue* to scores of *query sequence vs randomly shuffled homologue*. If the match is still high once the homologue has been shuffled then a large element of the score has come from a similarity in nucleotide or amino acid composition and the statistical significance of the identified homology is likely to be lower than its score suggests (anyone who has generated AT rich nucleotide sequences will be familiar with several large chloroplast genomes!). The conventional way to assess this is with z scores (Doolittle, 1981). After a large number of comparisons with randomly shuffled sequences of the same composition as the potential homologue, z is calculated as:

$$\frac{\text{(similarity score - mean of random similarity scores)}}{\text{standard deviation of random similarity scores}}$$

Depite these efforts, the investigator is still left with a statistic and not a probability, and must apply some subjective judgement to the final result. In general, a value of $z<3$ is not considered significant, while z values in excess of 6 are probably significant (Lipman and Pearson, 1985). There remains, however, an intervening grey area where results may or may not be significant, and the subjective judgement of the investigator may be required.

2.12 POLYMERASE CHAIN REACTION TECHNIQUES

The polymerase chain reaction (PCR) was used to amplify *Drosophila* genomic DNA and cDNA fragments and to generate single-stranded template for sequencing. Inverse PCR was used to screen cDNA libraries, a technique fully described in chapter 4. All PCR reactions were carried out in a Thermal Reactor (Hybaid).

2.12.1 PCRs to generate double stranded DNA products

PCRs yielding dsDNA products were performed using modifications of standard protocols (Innis, *et al.*, 1990). In each reaction, *Drosophila* genomic DNA or cDNA library template was present at 0.5-2.5ng/µl. Primer concentrations were usually 300nM; each dNTP was present at 100µM. Promega 10x Taq buffer was used to give working concentrations of 50mM KCl, 1.2mM $MgCl_2$ and 10mM tris-HCl pH 9.0 (at 25°C). Taq polymerase was included at 0.02-0.05 U/µl. Programs generally consisted of an initial 2 minute denaturation at 94°C, followed by 20-30 cycles of annealing-extension-denaturation. Annealing was done for 1 minute at a temperature between 1 and 3°C less than the lower of the predicted primer T_ms. Extension was at 72°C for 1 minute, or for 40 seconds/kb of target when amplifying large fragments. The denaturing step was at 94°C for 40-60 seconds.

2.12.2 Asymmetric PCR

Asymmetric PCR (Gyllensten and Ehrlich, 1988; McCabe, 1990) was used to generate single stranded template for sequencing. This was done either as a one step reaction, with the primer forming the 5' end of the single strand present at 800nM and the other primer present at 6nM, or as a two step reaction in which a 50 to 100µl double stranded PCR was done, run out on a TBE agarose gel, gel purified (§2.5.1), and used as template in a subsequent PCR with only one primer. The two step procedure, while more laborious, gave better results and was more often used. Aside from the bias in primer concentrations, one step asymmetric PCR conditions were identical to those described above (§2.12.1). In the second PCR of the two step procedure, 0.3 to 2.0 µg of template were used (in the absence of a second primer, amplification is linear rather than geometric, thus demanding a higher template concentration). In this reaction, 20 cycles were sufficient to generate ample ssDNA product for several sequencing reactions. The single primer was present at 360nM. Aliquots of asymmetric PCR products were run on agarose gels to provide an indication of the quality and yield of single stranded product. The remainder of each reaction was purified using a selective alcohol precipitation (Brow, 1990): one volume 4M NH_4OAc, pH 7.0 and two volumes isopropanol were added. After a 10 minute incubation at room temperature, DNA was recovered by spinning at 11000g for

10 minutes at room temperature, washed in 70% (v/v) ethanol, and air dried on the bench.

2.13 PHOTOGRAPHY AND AUTORADIOGRAPHY

Gel photographs were taken with a Polaroid Land Camera. Amersham Hyperfilm MP and Fuji RX film were used for autoradiography. ^{32}P signals were enhanced by exposure at -70°C with intensifying screens. Development was done either by hand, using D19 developer and Amfix fixer, or automatically in a Kodak X-Omat.

Three

One possible strategy for the isolation of molecules required for pathway and target selection in the developing insect nervous system is to screen for mutations which disrupt identified neuronal connections, and subsequently to clone and characterise the genes involved (§1.1.2). The giant fibre system of *Drosophila* is an ideal focus for such investigations. Its constituent neurons are large and identifiable (King and Wyman, 1980) and the system is required to mediate the escape response to a light-off stimulus, thus making it possible to isolate giant fibre system connectivity mutants as a subset of those mutagenised flies which fail to jump in response to a light-off stimulus (Thomas and Wyman, 1982; Thomas and Wyman, 1984b; Wyman and Thomas, 1983).

3.1 ANATOMY OF THE GIANT FIBRE SYSTEM

The anatomy of the giant fibre system is summarised schematically in Figure 3.1.

The cervical giant fibre (CGF) itself is a large interneuron whose cell body lies posteriorly in the lower protocerebrum of the brain. Giant commisural interneurons (GCIs) make electrical synapses with dendrites of both giant fibres within the brain (Phelan, *et al.*, 1996) so bridging the two cells[1]. The main CGF process runs anteriorly from the cell body and extends three branches before turning posteromedially and descending along the dorsal midline of the cervical connective into the thoracic ganglion. It continues its course posteriorly and ventrally into the mesothoracic neuromere where the axon makes a lateral bend and terminates (King and Wyman, 1980; Koto, *et al.*, 1981). The contralateral pair of CGF neurons appear to contact each other at a region of close apposition within the thoracic ganglion (King and Wyman, 1980). Although Koto and coworkers (Koto, *et al.*, 1981) have argued that the CGFs are not dye coupled in this region, recent experiments by Pauline Phelan and her colleagues (Phelan, *et al.*, 1996) have clearly indicated dye coupling between these fibres in the thoracic ganglion, presumably occuring at this identified region of close apposition. Similar coupling has also been demonstrated

1 At least one of the CGIs was previously mistaken by Koto and colleagues (Koto *et al.*, 1981) for part of the CGF dendritic tree.

between the CGF homologues in *Musca domestica* and *Calliphora erythrocephala* (Bacon and Strausfeld, 1986).

Within the mesothoracic neuromere of the thoracic ganglion, the CGF makes at least two further synapses (King and Wyman, 1980; Tanouye and Wyman, 1980). One of these is to the axon of the motor neuron driving the tergotrochanteral muscle (TTM, also called the tergal depressor of the trochanter, TDT), a muscle which extends the mesothoracic leg providing the main power for jumping (Nachtigall and Wilson, 1967) and is also likely to be responsible, indirectly, for wing elevation at the onset of flight (Tanouye and King, 1983). The cell body of the tergotrochanteral muscle motorneuron (TTMn) lies in the periphery of the thoracic ganglion, lateral to the point of synapse with the ipsilateral CGF (Baird, *et al.*, 1993). One TTMn process travels posteriorly from this region to make connections within the ganglion neuropil. The largest TTMn neurite runs from the point of synapse at the end of the CGF axon, laterally towards its cell body, to which it is connected by a short process. The large neurite then exits the thoracic ganglion via the posterior dorsal mesothoracic nerve (PDMN) (Baird, *et al.*, 1993; King and Wyman, 1980; Swain, *et al.*, 1990), and leaves the PDMN in a lateral branch to contact the ipsilateral TTM (King and Wyman, 1980). On the basis of lucifer yellow dye coupling (Koto, *et al.*, 1981; Phelan, *et al.*, 1996), cobalt coupling (Bacon and Strausfeld, 1986; Strausfeld and Obermeyer, 1976) and speed of impulse transmission (Tanouye and Wyman, 1980), it is believed that the CGF to TTMn synapse is electrical, i.e. a gap junction.

The CGF makes a further electrical synapse within the mesothoracic neuromere. This connection is to the peripherally synapsing interneuron (PSI) (King and Wyman, 1980). The PSI axon crosses the midline and exits the thoracic ganglion via the contralateral PDMN, but only travels about 20 µm into this nerve before making numerous reciprocal synapses with the dorsal longitudinal muscle motor neurons, which travel alongside it within the nerve, and then terminating (King and Wyman, 1980). Unlike the CGF synapses, these connections have morphological (King and Wyman, 1980) and physiological (Tanouye and King, 1983) characteristics of chemical synapses. In addition, the contralateral pair of PSIs have a large area of mutual contact at their decussation. Dye filling into one CGF invariably results in staining of both PSIs, indicating that each PSI synapses with both CGFs and/or the other PSI (Phelan, *et al.*, 1996). Each PSI also contacts its ipsilateral TTMn axon.

The five dorsal longitudinal muscle motor neurons (DLMns) serve the six fibres of the dorsal longitudinal flight muscle (Ikeda and Koenig, 1988; Ikeda, *et al.*, 1980) which powers the wing downstroke during flight (Pringle, 1949; Roeder, 1951). From the points of contact with the PSI within the PDMN, the large DLMn fibres travel in ribbon-like formation into the medial branch of the PDMN where they split up to innervate one (DLMns1-4) or two (DLMn5) muscle fibres (Ikeda, *et al.*, 1980). The positions of the somata and dendritic projections of the DLMn neurons have been elegantly mapped (Ikeda and Koenig, 1988). The cells bodies of DLMns 1-4 lie ipsilateral to the muscles they innervate, in the ventrolateral part of the thoracic ganglion, at the border of the pro- and meso-thoracic neuromeres. The DLMn5 cell body is a large profile lying contralateral to its muscle target, next to the midline in the outermost cell layer of the thoracic ganglion, at the border between the pro- and meso-thoracic neuromeres. Despite this difference in cell body positions, the dendritic trees of all the DLMn neurons project to similar regions of neuropil. Information regarding the morphology of some giant fibre system components is summarised schematically in Figure 3.1.

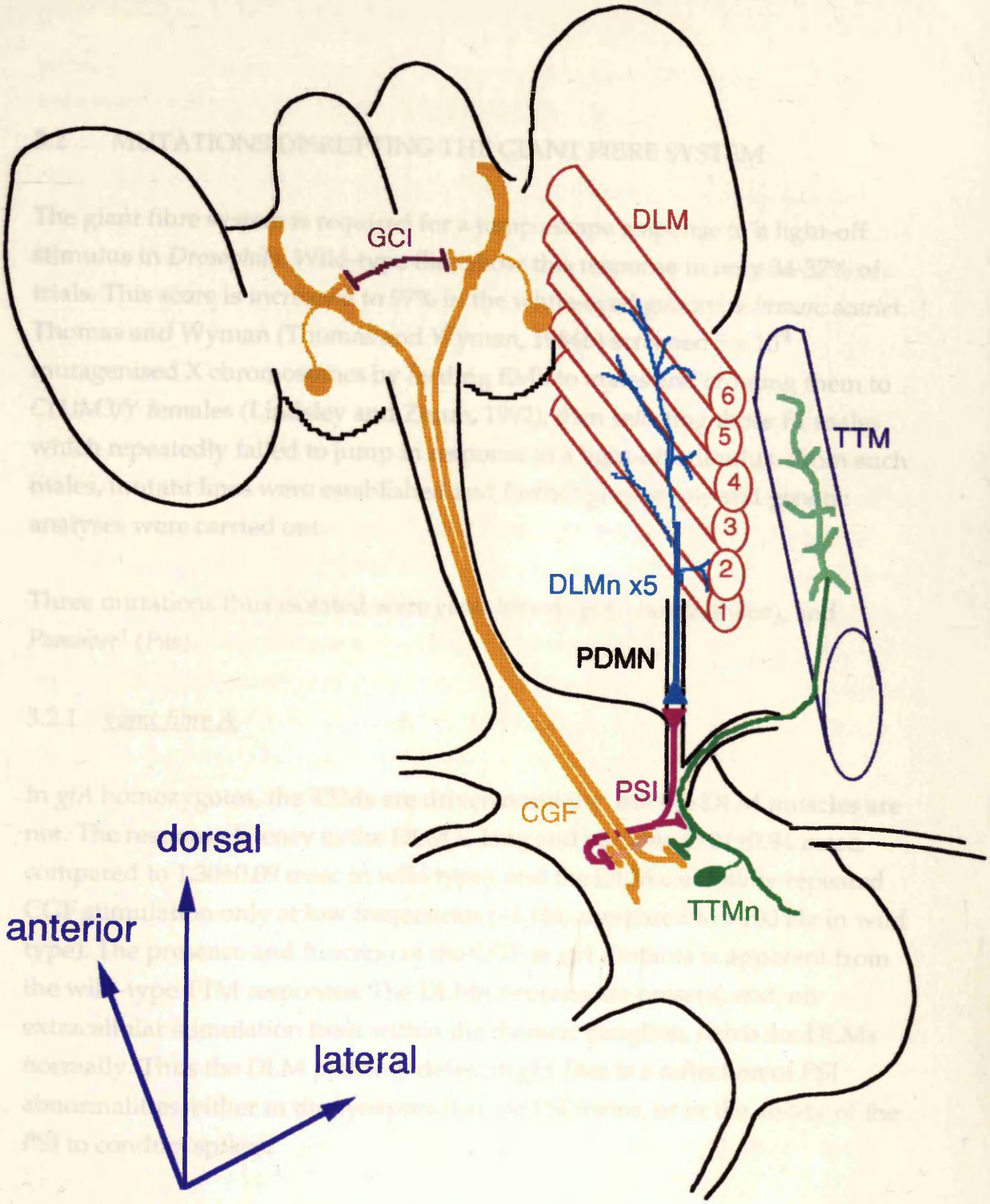


Figure 3.1. Schematic representation of the positions of, and interconnections among, neurons of the *Drosophila* giant fibre system. The black line represents the outline of the CNS. The CGF neurons are shown in amber. Both CGF neurons are indicated, while other components are shown for one side only. The PSI neuron is shown in pink, the DLMn in blue, the TTMn in green, and a GCI in purple. Multiple GCI neurons are known to exist; only one is shown. In the interests of clarity, the PSI, DLMn and GCI cell bodies and four of the five DLMn neurons are not shown. The DLM and TTM muscles are shown in red and blue respectively. Abutting parallel lines depict electrical synapses, while triangular terminals indicate chemical synapses. See text for further details.

3.2 MUTATIONS DISRUPTING THE GIANT FIBRE SYSTEM

The giant fibre system is required for a jump escape response to a light-off stimulus in *Drosophila*. Wild-type flies show this response in only 34-37% of trials. This score is increased to 97% in the white-eyed genotype *brown; scarlet*. Thomas and Wyman (Thomas and Wyman, 1984b) screened 5×10^4 mutagenised X chromosomes by feeding EMS to males and crossing them to *C(1)M3/Y* females (Lindsley and Zimm, 1992), then selecting those F1 males which repeatedly failed to jump in response to a light-off stimulus. From such males, mutant lines were established and further phenotypic and genetic analyses were carried out.

Three mutations thus isolated were *giant fibre A (gfA)*, *bendless* (*ben*), and *Passover*[1] (*Pas*).

3.2.1 *giant fibre A*

In *gfA* homozygotes, the TTMs are driven normally, but the DLM muscles are not. The response latency in the DLM is long and variable (3.21±0.94 msec, compared to 1.30±0.09 msec in wild type), and the DLM can follow repeated CGF stimulation only at low frequencies (~1 Hz, compared to >100 Hz in wild type). The presence and function of the CGF in *gfA* mutants is apparent from the wild-type TTM responses. The DLMn neurons are present, and, on extracellular stimulation from within the thoracic ganglion, drive the DLMs normally. Thus the DLM pathway defect in *gfA* flies is a reflection of PSI abnormalities, either in the synapses that the PSI forms, or in the ability of the PSI to conduct spikes.

All ipsilateral DLMn neurons are driven by the PSI, thus if the mutation caused conduction failure by the PSI, all ipsilateral DLM fibres would either fail or fire together after CGF stimulation. This is not found to be the case. Instead, individual ipsilateral fibres fire independently, implying that the *gfA* defect is at least partly a manifestation of abnormal PSI to DLMn synapses. *gfA* was mapped by recombination to a position proximal to *forked*, and was

---

1 This mutation was originally designated *nj156*, then *passover*. Baird *et al.*, (1990) demonstrated semidominant effects of this mutation, thus it is given a capital letter here.

subsequently shown to be uncovered by the deficiency *Df(1) JA27*, placing it between 18A5 and 18D1 on the cytogenetic map. It was found to be an allele of *lethal(1)myospheroid*, the locus encoding the *Drosophila* homolog of the β subunit of vertebrate integrins (Leptin, *et al.*, 1989). Integrins are heterodimeric cell surface receptors for components of the extracellular matrix, thus a gene with the potential to provide positional information to developing axons had been isolated, and the strategy underlying Wyman and Thomas' screen was vindicated.

3.2.2 *bendless*

In *bendless* homozygotes, the CGF axon fails to make its normal lateral bend just before the axon terminates in the mesothoracic neuromere of the thoracic ganglion (Thomas and Wyman, 1982). The morphology of the TTMn, however, appears normal, its medial process terminating at the normal point of contact with the CGF. The pathway to the dorsal longitudinal flight muscles has normal electrophysiological characteristics in *ben* flies, thus the CGF is present and neither the CGF to PSI nor the PSI to DLMn synapse is affected by the mutation. As would be expected, given the morphological abnormality of *bendless* giant fibres, the pathway to the TTM shows abnormal electrophysiology. The normal latency of the TTM response following extracellular stimulation of the CGF in the brain is 0.8 ms. In *ben* homozygotes this latency is increased to 2.2 to 3.0 ms. The ability of the TTM to follow repeated stimuli is also severely affected. While the TTM of wild type flies can be driven at frequencies greater than 100 Hz, the mutant only responds reliably to stimuli at frequencies of less than 1 Hz. Direct extracellular stimulation of the TTMn from the thoracic ganglion demonstrates a wild type latency and following frequency for muscle activation, thus neither TTMn conductivity, the TTMn to TTM neuromuscular junction, nor the jump muscle itself is affected by the mutation.

It is important to consider the relationship between the morphological abnormalities in *ben* flies and the electrophysiology of the CGF to TTM pathway. Broadly speaking there are two alternative explanations for the observed increases in latency and following frequency. Either the CGF to TTM synapse is defective but retains some vestiges of function, or the normal synapse is completely nonfunctional but an alternative neuronal pathway, with longer latency and lower ability to follow repeated stimuli remains active in the

mutant. Circumstantial evidence points to the latter of these possibilities. The increase in latency of the CGF to TTM pathway is in the order of 2.2 ms while a typical chemical synaptic delay is 0.5 ms, thus making it unlikely that a monosynaptic CGF to TTMn pathway is retained in the mutant (Thomas and Wyman, 1982). If the CGF to TTMn synapse is indeed nonfunctional, then the residual activity of the CGF to TTM pathway could result either from activation of the large TTMn by an alternative presynaptic pathway or from activation of one or both of the two small TTM motor neurons (sTTMn's) noticed in *Drosophila* by King and Wyman (1980), and positively identified as TTM motor neurons in *Calliphora* and in *Musca* by Bacon and Strausfeld (1986). In this context it would be valuable to find out whether the large TTMn is activated with a long latency by CGF stimulation in *bendless* mutants, though such an experiment is beset by technical difficulties. A rather easier question to answer would be whether the threshold CGF stimulation for long latency TTM activation in *ben* flies is identical to the threshold for activation of the normal TTM response in wild type. If this is not the case then it is likely that a different descending pathway (ie not the giant fibre) is responsible for the TTM response in the mutant, consistent with complete lack of function of the CGF to TTMn synapse. The question of relationships between morphological and electrophysiological abnormalities of flies carrying mutations affecting the giant fibre system is also relevant to the consideration of mutations at the *shaking-B* locus, and will be discussed further below.

Studies of gynandromorph mosaics have demonstrated that the *bendless* phenotype manifests itself when brain tissue (including the CGF neuron) is hemizygous for the mutation but not when the thoracic ganglion is mutant, thus implying that the *ben* gene product acts cell-autonomously within the CGF neuron (M.G. Muralidhar, R. Johnson and J. B. Thomas, personal communication). The *ben* phenotype is not a consequence of some nonspecific enfeeblement of CGF axon outgrowth. In the viable *bithorax* mutant genotype *abx* bx^3 *pbx/Df(3R)P2* the metathorax is transformed towards mesothoracic segmental identity. In this genotype the giant fibre displays its characteristic terminal bend in the mesothorax yet the axon trunk extends down into the transformed metathorax where it makes another lateral bend before terminating (Thomas and Wyman, 1984a). The interpretation of this result is that the CGF axon is responding to guidance cues which are duplicated in the homeotic mutant genotype. In flies which are *ben; abx* bx^3 *pbx/Df(3R)P2* the CGF fails to make lateral bends, yet still continues growth into the transformed

metathorax, thus demonstrating that the CGF axon in *ben* mutants retains a capacity for further outgrowth. The simplest explanation of these results is that the *ben* mutation disrupts the ability of the CGF axon to recognise signals which normally guide its terminal toward the region of synapse with the TTMn; in other words, according to the conceptual framework introduced in chapter 1, *ben* is a pathway selection mutation. Interestingly, the terminal region of the CGF axon is often observed to extend supernumerary processes in *ben* flies, which may reflect the fruitless search for a pathway which the cell is unable to recognise.

The CGF abnormality is the best characterised phenotype of *ben* homozygotes, but is not the sole defect. *ben* flies are also deficient in grooming (Phillis, *et al.*, 1993) suggesting neurological abnormalities outwith the giant fibre system.

ben was mapped by recombination to a region between *vermillion* and *forked*. The deficiency *Df(1)HA92* (Lindsley and Zimm, 1992) was found to uncover the mutation, placing *ben* in the 12A6-7 to 12D3 region (Wyman and Thomas, 1983). DNA covering the locus has now been cloned and the *ben* transcription unit has been identified and is homologous to the family of ubiqutin conjugating enzymes involved in targetting cellular products for protease degradation (Muralidhar and Thomas, 1993; Oh, *et al.*, 1994). The significance of this function with regard to axon pathfinding or the establishment of synapses is yet to be elucidated, but one intriguing possibility is that degradation of molecules mediating pathfinding signals directing posterior growth by the CGF must be degraded before 'lateral bend' signals may be responded to.

3.2.3 *Passover*

Unlike the situation in *bendless*, the CGF morphology in *Passover* homozygotes is wild-type. The TTMns are, however, abnormal: HRP backfills from the TTM reveal abnormal morphology of the medial neurite of the TTMn. In some of the first *Pas* TTMn neurons studied, the medial neurites were observed to "pass over" their normal point of contact with the CGF, cross the midline, and continue for some distance along the pathway of the contralateral TTMn. However, more rigorous study has revealed that this phenotype does not occur at a significantly higher frequency in *Passover* mutants than it does in some wild type control genotypes (Baird, *et al.*, 1993). The midline crossing phenotype originally ascribed to the *Passover* allele seems instead to be

attributable to the *Df(1)16-3-22* chromosome used in the original studies by Koto (M. Koto, (1983). Ph.D. thesis. Yale University). This chromosome is deficient for around 18 loci in the 19D1 to 20A2 polytene region (Schalet and Lefevre, 1976), including the locus of which *Passover* is an allele (see below). *Df(1)16-3-22* exerts a dominant midline-crossing phenotype, whose underlying genetic cause is not clear.

Athough midline crossing by the TTMn has been shown not to be a *Passover* mutant phenotype, the *Passover* mutation does affect both the length and the diameter of the TTMn medial neurite, (Baird, *et al.*, 1993; Swain, *et al.*, 1990). *Passover* mutants show a dosage-dependent reduction in both the medial and anterior extents of this projection, although other aspects of TTMn morphology are unaltered in the mutants. While this dendritic reduction has the effect of reducing the area of apposition of the TTMn and CGF neurons, it is important to note that reduced opportunity of contact between the CGF and TTMn is probably not the cause of the synaptic defects. In most cases the mutant TTMn medial dendrite is still observed to come within filopodial contact range[1] of the CGF, yet a stable synapse is never formed. Indeed, in some genotypes (e.g *Pas/Df(1)16-3-22*) the area of TTMn to CGF apposition is substantially greater than in wild type, yet the synapse is absent or abnormal. Thus it appears that it is not pathfinding *per se* that is defective in *Passover* flies. Instead synaptic target recognition may be disrupted. Alternatively synapse formation may be disrupted due to the lack of a protein required for the formation of the mature synapse.

The electrophysiology of TTM responses is also highly abnormal in *Pas* homozygotes. Brain stimulation elicits only a delayed (1.6 ± 0.8 ms SD) and intermittent TTM response which follows repeated stimuli only at frequencies less than 1 Hz (Baird, *et al.*, 1990; Thomas and Wyman, 1984b); i.e. a phenotype similar to, but distinct from, that of *ben* homozygotes. Extracellular stimulation of the TTMn's in the thoracic ganglion evokes wild type TTM responses, thus the TTMn's are present and able to transmit spikes, and both the TTMn to TTM neuromuscular junction and the jump muscle itself are unaffected by the mutation. Furthermore, the CGFs are present and able to conduct spikes in *Pas* flies (Thomas and Wyman, 1984b) thus implicating the CGF to TTMn synapse as the site of the TTM pathway defect. Because both *ben* and *Pas* homozygotes

---

1 This is typically around 30μm in the developing CNS (Goodman *et al.*, 1984) but may be 50-100μm for peripheral axon pioneers (Bentley and Keshishian, 1982).

appear to lack synaptic contact between the CGF and TTMn neurons, it may be that the TTM response to brain stimulation in these mutants reflects the electrophysiological characteristics of alternative brain to TTM pathways which are unmasked in the absence of a CGF to TTMn synapse. The observation that the *Pas* and *ben* TTM responses are distinct may reflect differential disruption of alternative brain to TTM pathways in these two mutants. However, the alternative explanation that the mutations impede but do not abolish CGF to TTM connectivity cannot absolutely be ruled out.

In *Pas* mutants, the pathway from the CGF to the DLM muscles is also disrupted. In homozygotes, brain stimulation never elicits a DLM response. The DLMns are still present and able to drive the DLMs normally if stimulated extracellularly in the thoracic ganglion (Thomas and Wyman, 1984b), thus the defect in the DLM pathway is likely to reside at either the CGF to PSI electrical synapse or the PSI to DLMn chemical synapse, or possibly both.

It would be surprising if any gene product were so specific in its action that mutants had defects in only one neuronal pathway, and flies homozygous for *Pas* or its allele *shak-B*^{2} (see below) do indeed have a range of other phenotypes. *shak-B*^{2} mutants have abnormal electroretinograms (Homyk, *et al.*, 1980), with greatly reduced or absent on-off transients and a diminished corneal negative component. These ERG components have all been shown to be a consequence of the activity of laminar neurons (Heisenberg, 1971), thus implying a defect in the laminar cells themselves, or in their synapses with the retinular neurons in *shak-B*^{2} mutants. These mutants also show extended female courtship duration (Kevin O'Dell, unpublished observations), gustatory defects (Balakrishnan and Rodrigues, 1991), uncoordinated motion under ether anaesthesia (Homyk, *et al.*, 1980; Miklos, *et al.*, 1987) and deficient grooming behaviour (Phillis, *et al.*, 1993). Thus *Pas* must also cause defects in nervous system components distinct from the giant fibre system.

3.3 GENETIC ANALYSIS OF *SHAKING-B*

Pas was originally mapped proximal to *forked* (Wyman and Thomas, 1983) and it was subsequently found that *Df(1) 16-3-22*, a deficiency chromosome which

lacks the 19D3 to 20A region (Schalet and Lefevre, 1976), fails to complement[1] *Pas*. Koto (M. Koto, (1983); PhD thesis, Yale University) subsequently refined the location of *Pas* to within *Df(1) B57*, a chromosome deficient for all of 19E but not for neighbouring polytene subdivisions (Schalet and Lefevre, 1976). Baird and coworkers (Baird, *et al.*, 1990) undertook a detailed genetic analysis of *Pas*. The cytological map position was first refined to the 19E2 to 19E6 region, as *Df(1) mal*^{10} (Schalet and Lefevre, 1976) which encroaches on the 19E region from the distal side, and *Df(1) Q539* which removes the proximal part of 19E, (Schalet and Lefevre, 1976) both fully complement *Pas* (a map showing the genetic extents of relevant deficiency and duplication chromosomes is shown in Figure 3.2).

Complementation tests with mutant alleles of loci lying between the *Df(1) mal*^{10} and *Df(1) Q539* breakpoints revealed that only alleles of the *R-9-29* complementation group[2] failed to complement *Pas*.

1 Note that because the *Pas* allele is only partially dominant, both mapping with deficiencies and complementation testing with point alleles are still possible.

2 Baird *et al.*, (1990) were unable to obtain alleles of *little fly* (19E6) (Schalet and Lefevre, 1976) which they could reliably score, hence this locus could not be tested.

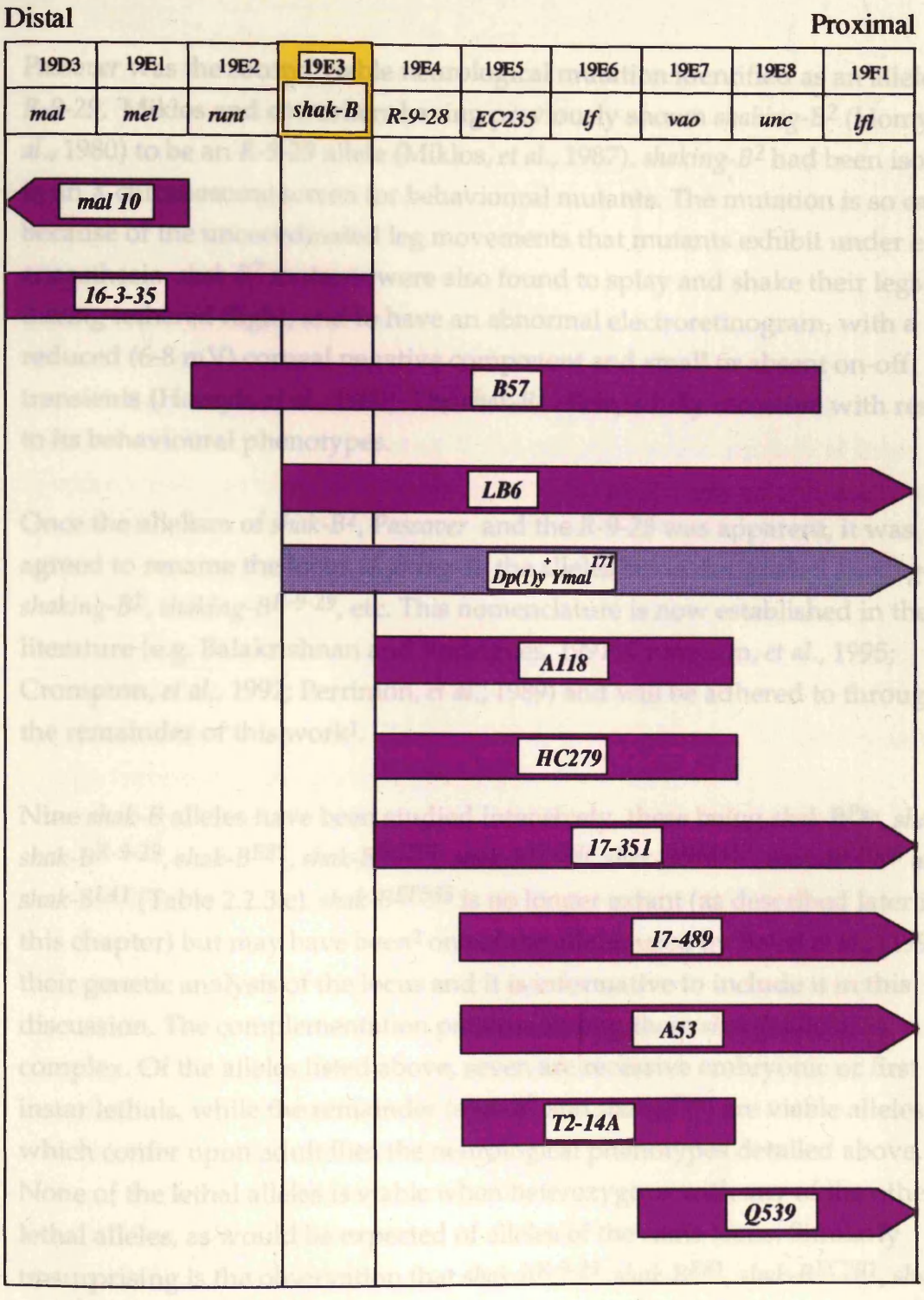


Figure 3.2. Genetic map showing deficiency and duplication chromosomes with breakpoints in the *shaking-B* region. Pink bars represent deficiencies. Mauve bars are duplications. Pointed ends indicate that the deficiency or duplication continues outwith the region shown.

Passover was the second viable neurological mutation identified as an allele of *R-9-29*, Miklos and coworkers having previously shown *shaking-B*^{2} (Homyk, *et al.*, 1980) to be an *R-9-29* allele (Miklos, *et al.*, 1987). *shaking-B*^{2} had been isolated in an X chromosome screen for behavioural mutants. The mutation is so called because of the uncoordinated leg movements that mutants exhibit under ether anaesthesia. *shak-B*^{2} mutants were also found to splay and shake their legs during tethered flight, and to have an abnormal electroretinogram, with a reduced (6-8 mV) corneal negative component and small or absent on-off transients (Homyk, *et al.*, 1980). The *shak-B*^{2} allele is fully recessive with respect to its behavioural phenotypes.

Once the allelism of *shak-B*^{2}, *Passover* and the *R-9-29* was apparent, it was agreed to rename the locus ***shaking-B***, the alleles being designated *shaking-B*^{Pas}, *shaking-B*^{2}, *shaking-B*$^{R\text{-}9\text{-}29}$, etc. This nomenclature is now established in the literature (e.g. Balakrishnan and Rodrigues, 1991; Crompton, *et al.*, 1995; Crompton, *et al.*, 1992; Perrimon, *et al.*, 1989) and will be adhered to throughout the remainder of this work[1].

Nine *shak-B* alleles have been studied intensively, these being *shak-B*^{Pas}, *shak-B*^{2}, *shak-B*$^{R\text{-}9\text{-}29}$, *shak-B*^{E81}, *shak-B*^{EC201}, *shak-B*$^{17\text{-}360}$, *shak-B*^{HM437}, *shak-B*^{EF535} and *shak-B*^{L41} (Table 2.2.3.c). *shak-B*^{EF535} is no longer extant (as described later in this chapter) but may have been[2] one of the alleles used by Baird *et al.*, (1990) in their genetic analysis of the locus and it is informative to include it in this discussion. The complementation patterns among these *shak-B* alleles is complex. Of the alleles listed above, seven are recessive embryonic or first instar lethals, while the remainder (*shak-B*^{2} and *shak-B*^{Pas}) are viable alleles which confer upon adult flies the neurological phenotypes detailed above. None of the lethal alleles is viable when heterozygous with any of the other lethal alleles, as would be expected of alleles of the same locus. Similarly unsurprising is the observation that *shak-B*$^{R\text{-}9\text{-}29}$, *shak-B*^{E81}, *shak-B*^{EC201}, *shak-B*$^{17\text{-}360}$ and *shak-B*^{HM437} all fail to complement the nervous system defects manifested by the viable alleles. These results in isolation hint that *shak-B*^{Pas} and *shak-B*^{2} could be viable alleles at a locus whose null phenotype is recessive lethality, and this was indeed speculated by Miklos and coworkers (Miklos, *et*

---

1 An attempt to simplify allele nomenclature still further has been made (Lindsley and Zimm, 1992). In this system the *shak-B* alleles are redesignated *shak-B*^{1} to *shak-B*^{25}. This system has not yet been adopted in other published work and in the interest of clarity will not be used here.

2 The reason for this uncertainty is explained later in this chapter.

al., 1987). Evidence that this is not, however, the case comes from studies of the *shak-B*^{EF535} and *shak-B*^{L41} alleles. Both of these lethal alleles fully complement the neurological mutations *shak-B*^{2} and *shak-B*^{Pas}, thus these mutations must be affecting distinct genetic functions. Furthermore, the neurological abnormalities of *shakB*^{2}*/Df(1)B57* are no more severe than those of a *shak-B*^{2} homozygote, suggesting that *shak-B*^{2} may be a null allele with respect to the neurological function that it disrupts. Thus it appears that *shak-B* is a complex locus at which at least two genetic functions reside, one of these being a vital function required for viability, another being required for normal development of the imaginal nervous system. The fact that most (5 out of 9) alleles disrupt *both* of these functions and that three of these five alleles were induced by EMS tends to suggest that the essential and neural products of the *shak-B* locus share some common coding regions.

These distinct genetic functions will here be referred to as *shaking-B(lethal)* and *shaking-B(neural)*. It must be stressed, however, that *shak-B(lethal)* may also have a role in the development in the imaginal nervous system, a possibility which will be further discussed below (§6.2.3.b).

3.4 CLONING OF *SHAKING-B* BY CHROMOSOME WALKING

Given the fascinating neuronal phenotypes of *shak-B* mutants, we wished to clone the locus and characterise its products in molecular detail. Chromosomal walking (Bender, *et al.*, 1983) was used to clone *shaking-B*, starting from entry points generated by microcloning (Pirrotta, 1986) of the 19E-F region (Miklos, *et al.*, 1988; J. A. Davies, unpublished data).

The deficiency chromosomes *Df(1) 16-3-35* and *Df(1) LB6* both fail to complement any *shak-B* mutations, thus at least some of the *shak-B* locus must be contained in the DNA absent from both of these chromosomes. The distal breakpoint of *Df(1) LB6* and the proximal breakpoint of *Df(1) 16-3-35* have been mapped onto one 200kb chromosomal walk, termed 952 (Alan Griffin, unpublished data). Only 15kb of genomic DNA lie between these breakpoints. Single copy restriction fragments from this area, and from the region immediately proximal to it[1] were used to screen embryonic cDNA libraries for

---

1 *Df(1)A118, Df(1) HC279* and *Df(1) 17-351* all have distal breakpoints proximal to the proximal *Df(1) 16-3-35* breakpoint (Baird *et al.*, 1990; Alan Griffin unpublished data). *Df(1) A118, Df(1)HC279* and *Df(1)17-351* all complement the lethality of *shak-B* alleles,

clones representing transcripts which might be derived from *shak-B*. $7x10^5$ recombinant clones from four different libraries (three embryonic, and one adult) were screened by J. A. Davies, and one positive clone, λKE2, was recovered from a 3-12 hr library constructed by L. Kauvar (Poole, *et al.*, 1985).

3.5 CHARACTERISATION OF KE2(1.8)

λKE2 has an insert size of 2.8kb. Eco RI digestion releases two fragments of 1.0 and 1.8 kb (data not shown). While the 1.8kb fragment hybridises discontinuously to genomic DNA of the 952 walk, the 1.0kb fragment does not hybridise, suggesting that the 1.0 and 1.8kb fragments might be unrelated, and that the 1.0kb fragment might not represent a region of *shak-B* transcript. The library from which KE2 was derived was generated by ligating Eco RI linkers onto double stranded cDNA before cloning into λgt10 (Poole, *et al.*, 1985). Such a procedure enables the cloning of cDNA concatemers. It was later concluded that λKE2 contained two unrelated insert fragments, and that only the 1.8kb fragment might be derived from a *shak-B* transcript. The 1.0kb fragment was discarded, and will not be discussed further.

3.5.1 Pattern of hybridisation of KE2(1.8) to cloned genomic DNA

λ phages containing inserts covering the entire 952 walk were digested with Eco RI, run out on a 0.8% agarose gel, and transferred to a reinforced nitrocellulose filter. A ^{32}P labelled probe made by random priming (§2.8.1) of the 1.8kb fragment of λKE2 (KE2(1.8)) was hybridised to this filter, in order to confirm that the cDNA was derived from the 952 walk region, and to establish, at low resolution, its splicing pattern. KE2(1.8) was shown to hybridise to λ phages 9405, 9403 and 94C11 (data not shown). To confirm and refine these data, the same probe was used in hybridisation experiments with blots of subclones of λ9405 and double digests of λ94C11. These experiments demonstrated hybridisation of KE2(1.8) to at least three discontinous regions, spanning much of the DNA between the *Df(1) 16-3-35* proximal breakpoint, and the distal breakpoint of *Df(1) LB6* (data not shown). Thus KE2(1.8) is a *bona*

---

but fail to complement their neuronal defects. Thus some DNA required for *shak-B* neuronal function(s) resides proximal to the *Df(1) 16-3-35* proximal breakpoint, inspiring the use of clones from this region in the search for *shak-B* cDNAs. See §4.6 for further details.

fide cDNA clone, (not a contaminating genomic DNA clone), and it is derived from a region in which we might expect to find *shaking-B* transcripts.

3.5.2 Sequencing of KE2(1.8)

In order to sequence KE2(1.8), the fragment was first subcloned into pBluescript II KS(+). It had been shown previously that KE2(1.8) contained one Hin dIII and one Bam HI site (J. A. Davies, unpublished data). These restriction sites subdivide the clone into two fragments of approximately 0.75kb and one of 0.25kb. These three fragments were subcloned into pBluescript II KS(+) and were sequenced on both strands using both T3 and T7 promoter primers, and custom-made primers complementary to the cDNA sequence (§2.10.3). The full annotated sequence is included in the Appendix to chapter 4.

3.5.3 Analysis of the KE2(1.8) sequence

3.5.3.a Orientation of the KE2(1.8) cDNA

The KE2(1.8) cDNA fragment is 1817 bases in length. At one end of the sequence there is a stretch of 17 A residues. In the construction of the library from which KE2 was obtained, first strand cDNA synthesis was primed with an oligo T primer to initiate reverse transcription from mRNA poly(A) tails, thus the oligo-A segment is likely to represent the most 3' region of transcript represented in the KE2(1.8) cDNA fragment. This orientation was subsequently confirmed by the analysis of other cDNAs from directionally cloned libraries (see below), and by the identification of a single G residue at the presumptive 5' end which is not encoded by the genome (see below) and is thought to represent a reverse-transcribed 7 methyl guanosine cap. To ascertain the direction of transcription on the chromosome, the 0.25kb Eco RI-Hin dIII fragment from the 5' end of the cDNA was labelled with ^{32}P and hybridised to a Southern blot of Eco RI-digested λ phages of the 952 walk. This fragment showed hybridisation only to the most proximal region identified by hybridisation with whole cDNA (data not shown). Thus KE2(1.8) is derived from an mRNA which is transcribed in the proximal to distal direction.

3.5.3.b Open reading frames in KE2(1.8)

Conceptual translation of KE2(1.8) in all three forward frames (Figure 3.3a) does not reveal a large segment of open reading frame (ORF). The sequence of KE2(1.8) was derived on both strands, on at least one strand of the corresponding genomic sequence (see below), and on at least one strand of several other cDNA clones which have regions of overlap with KE2(1.8) (see below and chapter 4). Thus the absence of a large ORF was not due to sequencing errors.

In the absence of a large ORF, four techniques were used to assess which, if any, of the small potential coding sequences might represent a genuine peptide product: (i) assessment of coding potential using the *Testcode* algorithm (§2.11.3.a.i), (ii) analysis of codon bias within potential coding regions (§2.11.3.a.ii.), (iii) analysis of the sequence context preceding each putative translation start, (iv) database searching to reveal homologues of potential peptide sequences (§2.11.3.b).

3.5.3.b.i *Testcode* analysis of KE2(1.8)

The result of *Testcode* analysis (§2.11.3.a.i) of the KE2(1.8) sequence is given in Figure 3.3b. Horizontal lines divide the *Testcode* window into three segments. Within the uppermost segment the *Testcode* prediction is that the window of sequence is coding. In the lower region the prediction is of noncoding sequence, while in the central region the program is unable to classify the sequence as coding or noncoding. Sequence windows are misclassified about 5% of the time by *Testcode* (Fickett, 1982).

Figure 3.3b demonstrates two major testcode peaks within KE2(1.8). By comparing this figure with Figure 3.3a, showing, at the same scale, the possible ORFs of KE2(1.8) it can be seen that the *Testcode* peaks approximately coincide with two ORFs, both in the second coding frame, which are highlighted in green.

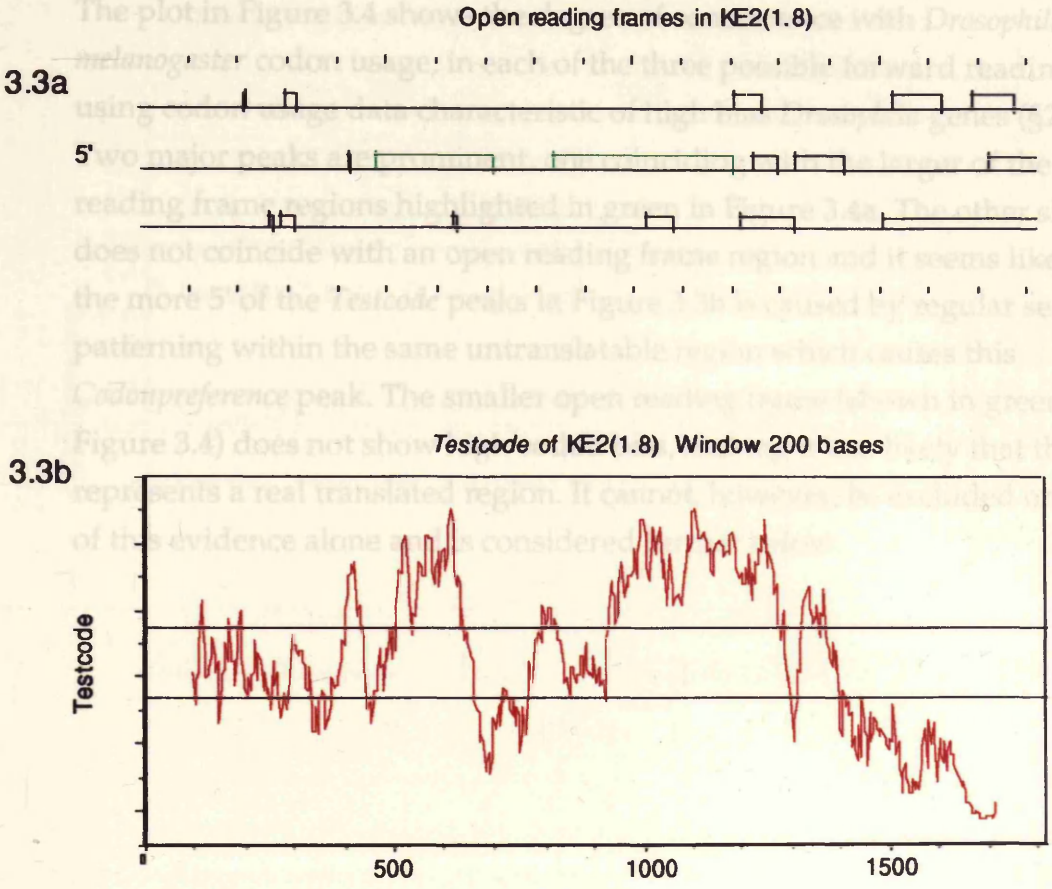


Figure 3.3: Assessment of open reading frames in the KE2(1.8) sequence. I: Identification of open reading frames and *Testcode* analysis.
3.3a: Boxes represent open regions of the three forward reading frames of KE2(1.8). Dashes above and below mark off the sequence into 100 base segments. The two largest frames are shown in green.
3.3b: *Testcode* analysis of the KE2(1.8) sequence. 2 distinct peaks are shown, corresponding to the regions of open reading frame shown in green in 3.3a, identifying these regions as good candidates for *bona fide* translated sequences.

3.5.3.b.ii *Codonpreference* analysis of KE2(1.8)

The plot in Figure 3.4 shows the degree of concurrence with *Drosophila melanogaster* codon usage, in each of the three possible forward reading frames, using codon usage data characteristic of high bias *Drosophila* genes (§2.11.3.a.ii). Two major peaks are prominent, one coinciding with the larger of the open reading frame regions highlighted in green in Figure 3.4a. The other sharp peak does not coincide with an open reading frame region and it seems likely that the more 5' of the *Testcode* peaks in Figure 3.3b is caused by regular sequence patterning within the same untranslatable region which causes this *Codonpreference* peak. The smaller open reading frame (shown in green in Figure 3.4) does not show high codon bias, making it less likely that this ORF represents a real translated region. It cannot, however, be excluded on the basis of this evidence alone and is considered further below.

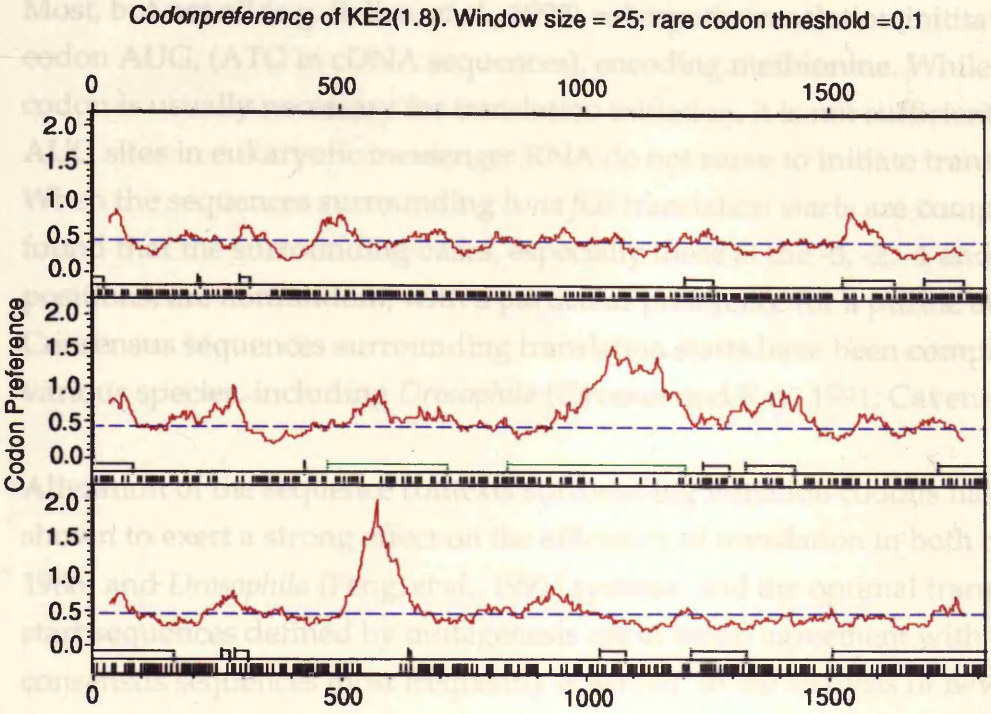


Figure 3.4: Assessment of open reading frames in KE2(1.8). II: *Codonpreference* analysis. The plots above show, in red, the degree to which windows 25 codons long throughout the KE2(1.8) sequence show codon usage which fits that observed in highly biased *Drosophila* genes (see §2.11.3.a.ii). The blue dotted lines show the average codon bias of a randomised sequence of the same base composition. The boxes over the baseline of each plot show open reading frames. The dashes below these mark the points at which rare codons occur. These are defined as those codons comprising less than 10% of the codons for that amino acid in the codon bias data set. Codon bias data are from Shields, *et al.*, (1988).

3.5.3.b.iii Potential translation starts in KE2(1.8)

Most, but not all (e.g. Bellen, *et al.*, 1992) eukaryotic translation inititates at the codon AUG, (ATG in cDNA sequences), encoding methionine. While an AUG codon is usually necessary for translation initiation, it is not sufficient, as many AUG sites in eukaryotic messenger RNA do not serve to initiate translation. When the sequences surrounding *bona fide* translation starts are compared, it is found that the surrounding bases, especially those at the -3, -2, -1 and +1 positions, are nonrandom, with a particular preference for a purine at -3. Consensus sequences surrounding translation starts have been compiled for various species, including *Drosophila* (Cavener and Ray, 1991; Cavener, 1987).

Alteration of the sequence contexts surrounding initiation codons has been shown to exert a strong effect on the efficiency of translation in both rat (Kozak, 1986) and *Drosophila* (Feng, *et al.*, 1991) systems, and the optimal translational start sequences defined by mutagenesis are in broad agreement with the consensus sequences most frequently observed. In the analysis of newly derived sequences, it is interesting to consider which of the possible AUG codons shares local sequence homology with identified translation starts, and as such might be a real initiation codon. It should be stressed, however, that there are many examples in the literature of biochemically confirmed translation starts which show poor homology to the consensus. This may be a reflection of the fact that consensus translation starts resemble optimal translation starts, while for some loci, suboptimal translation may be an important aspect of the control of gene expression. This said, however, there may be some sequences which the translational machinery is completely unable to recognise. Cavener (1987) found that the sequence YNNAUGY (where Y is a pyrimidine, and N is any base) was never found to occur within his sample of 100 *Drosophila* translation starts, while the expected frequency based on a null hypothesis of randomness of base composition surrounding the AUG codon is 25.

Cavener and Ray (1991) presented an analysis of 192 *Drosophila* AUG translation starts, showing the frequencies of occurrence of each base at different positions relative to the AUG codon. Given data in this form, assessment of goodness of fit of a new query sequence to the existing database can best be achieved by the use of a weight matrix (Staden, 1984). Weight matrix analysis basically involves converting the observed incidence of each

base at each position in a sample of aligned sequences into a measure of the probability of finding that particular base at that position (the probability weight matrix), by a suitable normalisation. Any new sequence is then scanned by a moving window, with the weight matrix probabilities for each position multiplied together to get a measure of goodness of fit to the original sample used to construct the matrix. In this case a weight matrix may be assembled by dividing the observed frequency of each base at each position by the expected random occurrence of that base in *Drosophila* nuclear genomic DNA, an estimate of which has been derived by Laird and McCarthy (Laird and McCarthy, 1968). When this normalisation is applied to the data of Cavener and Ray, the following weight matrix results:

	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	ATG	+4	+5
A	1.088	1.333	0.192	1.193	1.123	0.772	0.737	2.281	1.649	1.368		0.912	0.982
C	1.581	0.791	1.581	1.349	0.791	1.674	2.465	0.326	1.070	1.581		0.884	1.395
G	0.884	0.930	0.744	0.744	1.302	0.884	0.558	0.930	0.558	0.884		1.628	0.977
T	0.596	0.877	0.877	0.737	0.807	0.772	0.491	0.281	0.632	0.281		0.702	0.702

Table 3.1 Probability weight matrix used to generate the *Drosophila* translatability score T_{Dros}. Scores are based on the data of Cavener and Ray (1991), and Laird and McCarthy (1968).

A relative score of goodness of fit of a query sequence to the database of translation starts can then be achieved by multiplying together the scores that each base at each position in the query sequence is awarded in the probability weight matrix, and taking the natural logarithm of the resulting product. We can call this the *Drosophila* translatability score (T_{Dros}), and can use it to assess the goodness of fit of proposed translation starts to those investigated by Cavener and Ray.

Table 3.2 lists the ATG codons present in KE2(1.8), and shows the local sequence context in which each occurs, followed by the T_{Dros} score and the length of reading frame which would be initiated. All those ATG codons present within the context YNNATGY are marked with an x. It is unlikely that these reading frames can be translated efficiently, and each would, in any case, initiate only a tiny reading frame.

Position in KE2(1.8)	Local sequence context	T_{Dros} score	ORF length (codons)
211	CGATCGAAAA**ATG**TG	-0.972	1
261	AGCTTCGACC**ATG**TC	1.211	4
285x	ACCCGCGCTT**ATG**CC	-1.837	10
295	ATGCCGCGCG**ATG**AA	0.263	8
425x	AAAAGTGCCA**ATG**TA	-2.792	1
476	GGACCCGAAG**ATG**CC	-0.438	80
633x	CGATTAACCG**ATG**CG	-3.671	3
644 (476)	TGCGTCGATA**ATG**GG	0.433	24
839	ACGACAAACC**ATG**TT	-0.425	122
935 (839)	AATAACTGTG**ATG**AT	-0.764	90
944 (839)	GATGATTCTG**ATG**GT	-2.686	87
1026	ATTCCAGAGG**ATG**TG	-1.215	19
1207	AGCCGATATA**ATG**CG	0.596	19
1218x	TGCGGAACAG**ATG**TT	-2.179	17
1244x	TTGCAGCCAA**ATG**CC	0.152	18
1256 (1244)	GCCAACGACA**ATG**GC	2.350	14
1328	ACGGAATCGA**ATG**AT	-3.137	34
1340 (1328)	GATCGCCGAG**ATG**CA	2.176	30
1509	CAAGATCACG**ATG**AT	-0.118	100 +
1512 (1509)	GATCACGATG**ATG**GG	1.087	99 +
1528	CGACATTGTC**ATG**TG	-2.270	35
1693	ACAACTTATA**ATG**AG	-2.103	28
1724	ATACACCAAG**ATG**GC	2.157	27 +
1759 (1693)x	GGCAAGATTC**ATG**CG	-1.291	6
1770 (1509)	TGCGCTCATC**ATG**GA	1.274	12 +

Table 3.2 Translatability scores of potential open reading frames of KE2(1.8). Where a translation start is also a continuation of a reading frame already open, the most N terminal start of the same ORF is indicated in parentheses. Reading frames still open at the end of the KE2(1.8) sequence are marked with a "+".

Of the many potential ORFs within KE2(1.8), only one is predicted by *Testcode* and *Codonpreference* analysis to be a genuine coding sequence. This is the 122 residue ORF starting at position 839. While its translation start does not show a sequence context that is optimal for translation, it yields a T_{Dros} score that is much higher than those of some of the ATG codons present and it is not unreasonable that it might be a *bona fide* start site. If this reading frame region is

genuinely translated then, in common with approximately 5-10% of other eukaryotic mRNAs (Kozak, 1987) it will, at least in this splice form, have ATG codons upstream which may be involved in translational regulation (e.g. Lohmer, *et al.*, 1993).

The identification of this small candidate ORF by no means proves that it is genuinely the product of the mRNA represented by the KE2(1.8) cDNA. Furthermore, it was possible that all of KE2(1.8) was an untranslated region of a larger transcript. Further data were therefore needed to add legitimacy to this potential protein.

3.5.3.b.iv Homologues of the potential products of KE2(1.8)

Tfasta (Pearson and Lipman, 1988) searches with the 122 residue ORF revealed significant homologies with the products of two previously identified loci: the *Drosophila lethal (1) ogre* locus (Watanabe and Kankel, 1990), Fig. 3.5a and the *Caenorhabditis elegans* locus *unc-7* (Starich, *et al.*, 1993), Fig. 3.5b.

The anticipated product of the 122 amino acid ORF is a basic (estimated pI=9.2) 14.1 kDa protein. In this chapter I will present evidence that KE2 does indeed represent a transcript from the *shaking-B* locus, and will refer to this 14.1 kDa protein as Shak-B(14.1). The homologies of Ogre and Unc-7 with Shak-B(14.1) are highly significant: *rdf2* scores (Pearson and Lipman, 1988); (see also §2.11.3.b) are 36.42 and 12.35 respectively (recall that scores of 6 or more are considered highly significant).

Shak-B(14.1) shows 47.5% identity to the first 121 amino acids of the Ogre protein, with 1 gap. When conservative changes are accepted, the homology increases to 62.3%. The homology with the Unc-7 protein shows 28.3% identity and 39.1% similarity when three gaps are allowed:

(a)

```
Shak-B(14.1) 1   MLDIFRGLKNLVKVSHVKTDSIVFRLHYSITVMILMSFSLIITTRQYVGN
                 |  :   ||  :|   : ||  ||||| | | ::|:: ||||| ||||:
Ogre         1   MYKLLGSLKSYLKWQDIQTDNAVFRLHNSFTTVLLLTCSLIITATQYVGQ

             51  PIDCVHTKDIPEDVLNTYCWIQSTYTLKSLFLKKQGVSVPYPGIGNSDGD
                 || |:    :|  |:||:||| ||:|:  | :  |  |  ||::|  ||
             51  PISCI.VNGVPPHVVNTFCWIHSTFTMPDAFRRQVGREVAHPGVANDFGD

             101 PADKKHYKYYQWVCFCLFFQPI 122
                    || | ||||||| |||| :
             100 EDAKKYYTYYQWVCFVLFFQAM 121
```

(b)

```
Shak-B(14.1) 2    LDIFRGLKNLVKVSHVKTDSIVFRLHYSITVMILMSFSLIITTRQYVGNP
                  : ::        :     |  | :| |  |  || || |::: :|||| |
Unc-7       121   MILYYLASAFRALYPRLDDDFVDKLNYYYTTTILASFALLVSAKQYVG.P

             52   IDC....VHTKDIPEDVLNTYCWIQSTYTLKSLFLKKQGVSVPYPGIGNS
                  | |        |  |     |||:| ||             ||
            170   IQCWVPATTFTDAMEQYTENYCWVQNTY.............WVPMQEDIPR

             98   DGDPADKKHYKYYQWVCFCLFFQPI 122
                  :      :   ||||| | |    :
            208   EIYSRRNRQIGYYQWVPFILAIEAL 232
```

Figure 3.5 Homology between Shak-B(14.1) and (a) the N terminus of the Ogre protein; (b) an internal region of the Unc-7 protein. Identities are shown with lines (|), conservative changes (A,G; S,T;D,E;K,R;V,I,M,L;N,Q;Y,W,F) are indicated with colons (:).

When the KE2(1.8) sequence is used for DNA level searches, only homologies to the *l(1)ogre* and *unc-7* DNA sequences are observed. Further protein level searches were also done using nonsense proteins derived from translating KE2 in each possible frame and ignoring stop codons. Conceptual translation of a small 3' region of KE2 was found to yield a product with further homologies to *Ogre* and *Unc-7* proteins, as will be discussed in the next chapter. Translation of the remaining regions of KE2(1.8) fails to yield proteins with significant homologues in the databases, consistent with (though by no means proof of) the proposition that other regions of the KE2(1.8) sequence do not encode protein products. The homologies with *l(1)ogre* and *unc-7* will be discussed further in chapters 4 and 5.

3.5.4 Genomic organisation of the transcript represented by KE2(1.8)

In order to determine, at high resolution, the splicing pattern of KE2(1.8), genomic fragments containing all of the KE2(1.8) exon sequences, previously identified by hybridisation (see above), were subcloned into pBluescriptII™, restriction mapped with Eco RI, Eco RV, Hin dIII, Hin cII, Bam HI, Acc I, Pst I, Kpn I, Bgl II, Pvu II, Xba I, Xho I, Sal I, Sma I and Sst I, and fully or partially sequenced. A map showing the subcloned regions and the organisation of KE2(1.8) is given in Figure 3.6. The sequences derived from these subclones are included in the Appendix to chapter 4.

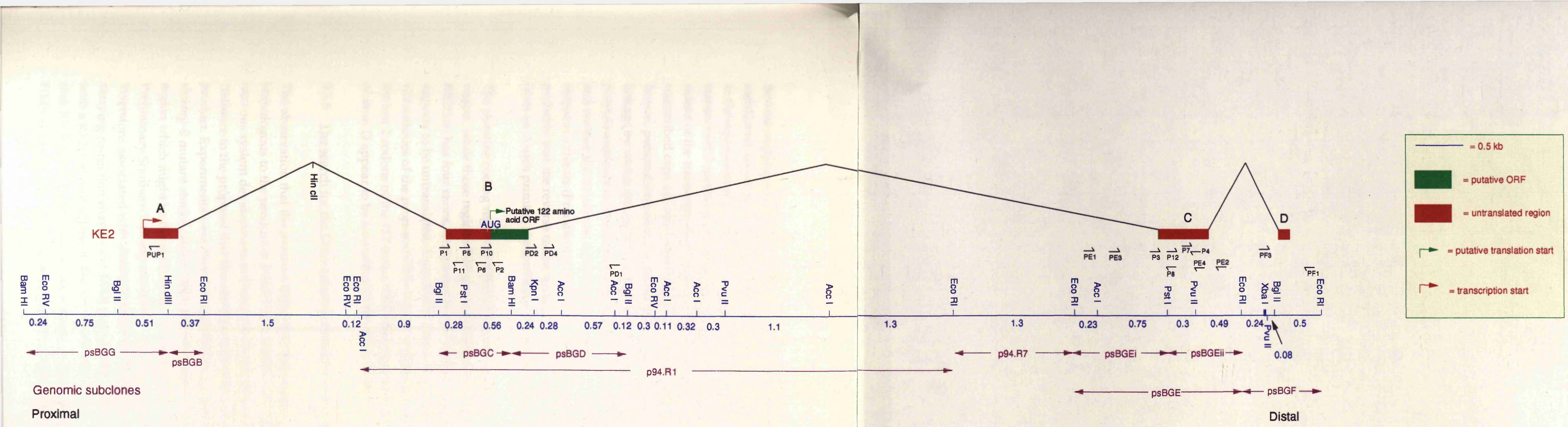


Figure 3.6: Genomic organisation of the transcript represented by the KE2(1.8) cDNA. Exons are shown as filled boxes. The putative untranslated regions are shown in red, the 122 codon ORF is shown in green. Primers used in the sequencing are shown underneath the exon boxes. A restriction map of the corresponding genomic DNA is shown below the exons. All named genomic subclones except p94.R7 have been sequenced. psBGB, psBGC, and psBGD have been completely sequenced; only partial sequences of the other genomic subclones have been derived. These sequences may all be found in the appendix to chapter 4.

Several important conclusions may be drawn from this analysis. Firstly, as mentioned above, the first base of KE2(1.8) is a G residue which is not encoded in the equivalent genomic position. This is likely to represent a reverse-transcribed 7-methyl guanosine cap, implying that KE2(1.8) includes the full 5' extent of the transcript from which it is derived. Remarkably, reverse transcribed caps are often found at the 5' ends of full length cDNA clones (Nick Brown, personal communication), in spite of the unusual 5' to 5' trisphosphate linkage by which the cap is attached. At the 3' end of KE2(1.8), the small poly(A) sequence is found to correspond to an A-rich sequence in the genome, and neither the consensus AATAAA polyadenylation signal, nor any of its frequent variants (Proudfoot and Whitelaw, 1988) is present. It is therefore probable that the reverse transcription event that yielded KE2(1.8) was initiated from an A-rich priming site internal to the mRNA molecule.

The putative coding region of KE2(1.8) is shown in Figure 3.6 as a filled green region, while those regions thought to be untranslated are shown in red. KE2(1.8) has four exons, designated A to D. Exon A comprises 375 bases and appears to be untranslated. Exon B is 823 bases long, and contains 120 of the 122 residues of the proposed ORF. Exon C is 504 bases in length and contains the last 2 codons of the 122 codon ORF before a stop codon, while the 115 bases of exon D appear to be noncoding in this splice form.

3.5.5 The search for the DNA lesions underlying *shak-B* alleles: Part I

The observation that the putative 122 amino acid product of KE2(1.8) is highly homologous to the product of *lethal(1)ogre*, which is involved in imaginal nervous system development in *Drosophila* (Watanabe and Kankel, 1990) lent credence to the proposal that this small protein represents a genuine *shaking-B* product. Experiments were therefore undertaken to determine whether any *shaking-B* mutant stocks carried DNA polymorphisms within the KE2(1.8) region which might be the underlying causes of *shak-B* mutant alleles. Preliminary Southern blot experiments to detect large insertions, deletions or inversions associated with *shaking-B* alleles were done by hybridising filters carrying restriction enzyme-cut DNA from various *shaking-B* mutant stocks with a KE2(1.8) probe (Jane A. Davies, unpublished data). Only the *shak-B*^{HM437}, *shak-B*$^{17\text{-}360}$ and *shak-B*^{L41} stocks were tested, as these alleles were induced with HMS, neutrons and X-rays respectively, all of which are mutagens which tend

to induce large deletions and rearrangements (Kramers, *et al.*, 1983; Pastink, *et al.*, 1987). No DNA polymorphisms in the *shak-B*^{HM437}, *shak-B*$^{17-360}$ or *shak-B*^{L41} stocks were revealed in these experiments (Jane A. Davies, unpublished results).

3.5.5.a The search for *shak-B* mutant lesions: Strategy

A high resolution search for DNA changes associated with *shaking-B* alleles was therefore undertaken. A substantial selection of techniques which allow detection of DNA polymorphisms down to the level of single base changes is now available, each having its associated strengths and weaknesses. These techniques include chemical cleavage to detect mismatches between mutant and wild-type sequences (Montadon, *et al.*, 1989), denaturing gradient (Myers, *et al.*, 1985) and temperature gradient gel analyses and single-stranded conformational polymorphism (SSCP) analysis. However in the context of looking for *shaking-B* mutant lesions, all of these techniques share a common problem: namely that, while they will detect polymorphisms with varying degrees of sensitivity, they cannot distinguish between neutral changes and lesions which underlie mutations. Before beginning the search for *shaking-B* mutant alleles it was apparent that some neutral polymorphisms would be present. For example, the BamHI site contained within the 122 codon ORF of KE2(1.8) is present in Oregon-R, but absent from Canton-S wild type DNA[1]. Because encounters with neutral polymorphisms were anticipated, and because the proposed coding regions of KE2(1.8) are relatively small, a strategy of directly sequencing PCR products generated from *shak-B* mutant DNA, was adopted.

Initial attempts to directly sequence large quantities (approximately 1.5 pmol) of double stranded PCR products yielded unsatisfactory results, probably because reannealing of the template strands is thermodynamically favoured over annealing of the sequencing primer. An asymmetric PCR (Gyllensten and Ehrlich, 1988; McCabe, 1990) strategy was therefore adopted to generate single stranded sequencing templates from *shaking-B* mutant alleles (§2.12.2). This approach has two major advantages over the alternative of sequencing cloned double stranded PCR products. Firstly, it enables the sequencing of DNA

---

1 While the sequence of this region has not been determined directly from Canton-S DNA, a single A vs G polymorphism (GGATCC vs GGGTCC) is observed on sequencing multiple *shak-B* cDNAs (see below), and it is probable that the GGGTCC variant is present in Canton-S.

mixtures derived from both mutant and balancer chromosomes without requiring the analysis of multiple isolates. Secondly, spurious polymorphisms derived from PCR errors are unlikely to be detected, as the sequencing template is a large population of PCR product molecules and Taq polymerase errors are unlikely to be detectable against the background of faithfully replicated template molecules.

Because most *shaking-B* alleles are embryonic or early first instar lethals, it is not possible to harvest DNA from the mutant chromosome in a homozygous or hemizygous state. DNA samples of lethal *shaking-B* alleles were therefore obtained from female flies carrying both the mutant allele and a balancer X chromosome. Point polymorphisms were detected as double bands on sequencing gels while deletions or insertions resulted in the superimposition of two sequence ladders, both of which are easily recognised. While highly sensitive for the detection of heterozygous polymorphisms, this PCR technique has a major limitation. If either or both PCR primers lie in a region which shows polymorphism in the mutant chromosome, then the resulting product may be a wild type fragment derived from the balancer chromosome alone (see also §6.1.1).

One attempt to circumvent this problem relied on the ability of closely related *Drosophila* species to mate to give viable (though sterile) progeny. If, for example, a male *D. simulans* and a female *D. melanogaster* mate, their only viable progeny will be interspecific hybrid females (Hutter and Ashburner, 1987; Hutter, *et al.*, 1990; Sturtevant, 1920). While *D. simulans* and *D. melanogaster* are closely related, they have diverged sufficiently to show substantial DNA polymorphisms in regions not tightly constrained by selection (Bodmer and Ashburner, 1984). It is therefore possible in theory to design a primer pair which will amplify a product from *D. melanogaster* but not *D. simulans* DNA and enable the distinction between a wild type sequence and a complete deficiency in the *D. melanogaster* mutant chromosome. A selection of primers was tested on *D. simulans* DNA in the hope of finding a pair that would *not* work, but all primer pairs insisted on doing so (data not shown). Thus, in the absence of detailed local sequence information from *D. simulans*, this technique could not be used.

While a wild type sequence would result from either a wild type region or a deleted region on the mutant chromosome, the detection of a heterozygous

polymorphism, regardless of its nature, implies that two sequences have been amplified. In the results presented below, I will include each heterozygous polymorphism detected, as this implies that the mutant chromosome does not carry a deletion in the region analysed. There remain cases where no heterozygous polymorphisms have been detected. In many such cases the possibility that the mutant chromosome carries a deletion of the region in question could be ruled out or confirmed by repeating the experiment using *D. simulans* x *D. melanogaster* hybrid females, in order to maximise the local sequence heterozygosity. These experiments have not yet been attempted.

Any observed DNA polymorphism can only be a candidate for a mutant lesion if that same polymorphism is not found in the progenitor chromosome in which the allele was induced. In every case, either the progenitor chomosome or a different allele from the same mutagenesis was available as a control:

shaking-B Allele	Control
shaking-B^{2}	Oregon-R
shaking-B^{Passover}	Canton-S
shak-B$^{R\text{-}9\text{-}29}$	*R-9-28*
shak-B^{E81}	Qa
shak-B^{EC201}	*EC235*
shak-B$^{17\text{-}360}$	*Df(1)17-489*
shak-B^{HM437}	*runt*^{HM449}
shak-B^{L41}	*M56i*
shak-B^{EF535}	*M56i*

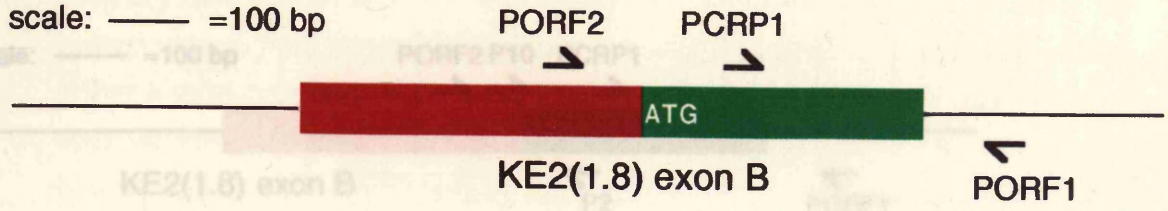
Table 3.3. Controls used to assess the significance of polymorphisms detected in DNA from *shak-B* alleles.

3.5.5.b The search for *shak-B* mutant lesions: The *shak-B*^{L41} lesion

Genomic DNA was prepared (§2.7.3) from the alleles listed in Table 3.3. PCR primers PORF1 and PORF2 (§2.10.3) were designed to allow amplification of a 542 base pair, double stranded fragment containing the 120 codons of the Shak-B(14.1) ORF residing in exon B of KE2(1.8), as shown in Figure 3.7. PCR was performed on DNA from *shak-B* alleles and appropriate controls using PORF1

and PORF2, and the resulting products were run out on a 30cm, 1.2% agarose gel. After running for 24 hours at 1.6V/cm, PCR samples from two *shak-B* alleles (*shak-B*^{L41} and *shak-B*^{EF535}) revealed double bands representing size polymorphisms between amplification products from mutant and balancer chromosomes (Figure 3.7). In order to investigate the nature of these polymorphisms and to look for any other DNA changes not detectable by agarose gel electrophoresis, all PCR products derived from *shak-B* mutant alelles were sequenced. This was achieved by first synthesising asymmetric PCR templates using each PCR primer individually. Single strands in both directions were synthesised for each double stranded fragment, and these were then sequenced using complementary PCR primers and complementary internal sequencing primers until the whole fragment had been sequenced on at least one strand (Figure 3.8).

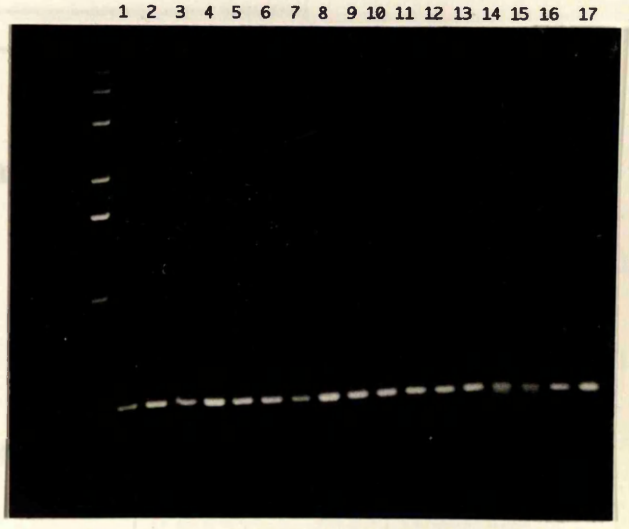
(a)



(b)

PCR between primers PORF1 and PORF2 was performed on the following genotypes:

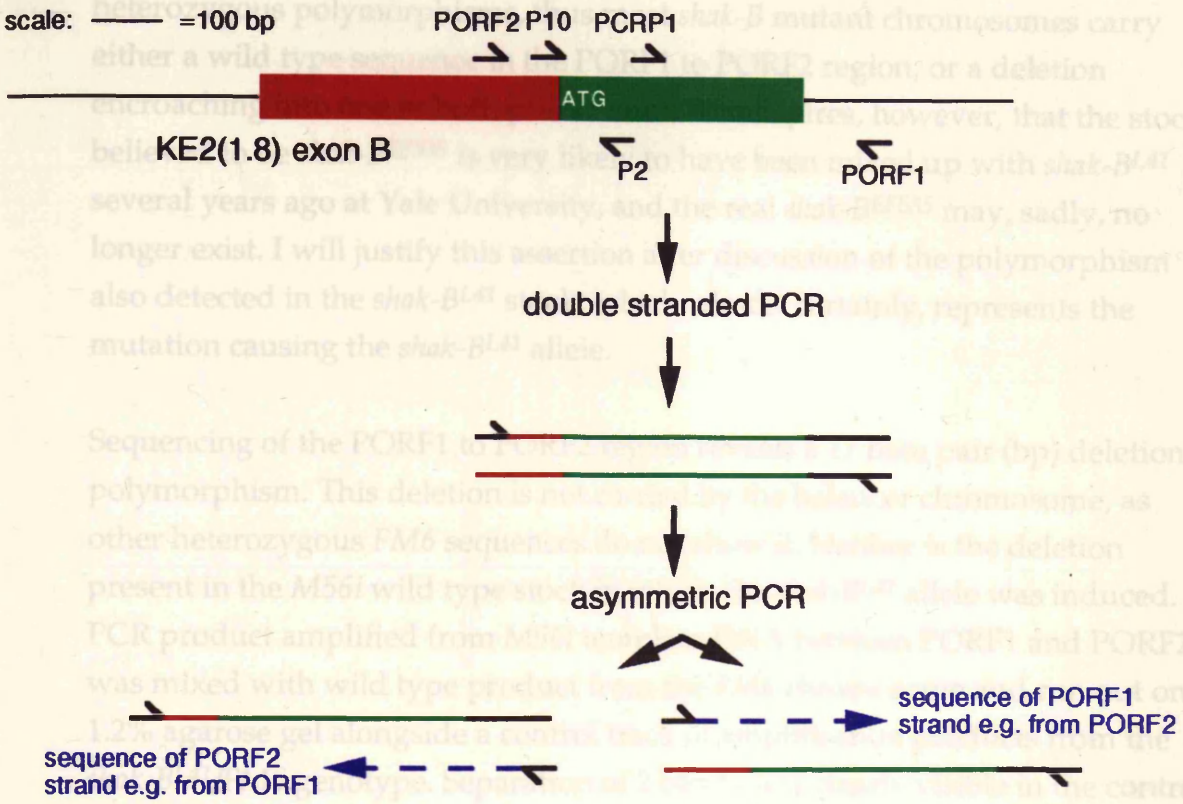
1. *shak-B^{R929}/FM6*
2. *R928/FM6*
3. *shak-B^{EC201}/FM6*
4. *EC235/FM7*
5. *shak-B^{E81}/FM6*
6. *Qa*
7. *shak-B^{HM437}/FM6*
8. *runt^{HM449}/FM6*
9. *shak-B$^{17\text{-}3\text{-}60}$/Binsn*
10. *Oregon-R*
11. *FM6*
12. *FM7*
13. *FM6/y^{+}Ymal106*
14. *"shak-B^{EF535}"/FM7*
15. *shak-B^{L41}/FM6*
16. *shak-B^{Passover}*
17. *shak-B^{2}*



A small deletion polymorphism is detectable in *shak-B^{L41}/ FM6* and *"shak-B^{EF535}"/FM7* genotypes between PORF1 and PORF2 primers. No polymorphism is detected between PORF1 and PCRP1 (data not shown) implying that in both genotypes a deletion exists in the region between PORF2 and PCRP1. Sequence analysis of PORF1 to PORF2 asymmetric PCR templates from all *shak-B* alleles was then undertaken to determine the nature of any DNA polymorphisms present (Figure 3.8).

Figure 3.7: Strategy for the detection of mutant lesions in the region of the Shak-B(14.1) ORF encoded by exon B of KE2(1.8). (a):diagram of KE2(1.8) exon B, showing primers used to amplify this region. Left is 5' and proximal, right is 3' and distal. The region encoding most of Shak-B(14.1) is shown in green, the small potential ORF of KE2(1.8) is not shown. (b): Results of double-stranded PCR amplification between PORF1 and PORF2 using *shaking-B* mutant and control DNAs as PCR templates. Quotation marks around *shak-B^{EF535}* reflect the questionable nature of this stock (see text).

(a)



(b) Summary of genotypes tested for mutations by PCR between PORF1 and PORF2:

Genotype	Sequence	Heterozygous Polymorphisms
shak-B^{R929}/FM6	wild type	none detected
shak-B^{EC201}/FM6	wild type	none detected
shak-B^{E81}/FM6	wild type	none detected
shak-B^{HM437}/FM6	wild type	none detected
shak-B$^{17\text{-}3\text{-}60}$/Binsn	wild type	none detected
shak-B^{L41}/FM6	17 bp deletion	mutant lesion
"shak-B^{EF535}"/FM7	17 bp deletion	mutant lesion
shak-B^{L41}/FM6;bw;st	17 bp deletion	mutant lesion
"shak-B^{EF535}"/FM6;bw;st	17 bp deletion	mutant lesion
shak-B^{Passover}	wild type	hemizygous template
shak-B^{2}	wild type	homozygous template

Figure 3.8 Asymmetric PCR analysis of PORF1 to PORF2 region in *shak-B* alleles. (a): Schematic diagram of asymmetric PCR method. (b):Summary of results from sequence analysis of asymmetric PCR templates from PORF1-PORF2 region. See text for details.

In every case except *shak-B*^{L41} and *shak-B*^{EF535}, sequence analysis revealed no heterozygous polymorphisms, thus most *shak-B* mutant chromosomes carry either a wild type sequence in the PORF1 to PORF2 region, or a deletion encroaching into one or both primer sites. It transpires, however, that the stock believed to be *shak-B*^{EF535} is very likely to have been mixed up with *shak-B*^{L41} several years ago at Yale University, and the real *shak-B*^{EF535} may, sadly, no longer exist. I will justify this assertion after discussion of the polymorphism also detected in the *shak-B*^{L41} stock, which, almost certainly, represents the mutation causing the *shak-B*^{L41} allele.

Sequencing of the PORF1 to PORF2 region reveals a 17 base pair (bp) deletion polymorphism. This deletion is not carried by the balancer chromosome, as other heterozygous *FM6* sequences do not show it. Neither is the deletion present in the *M56i* wild type stock in which the *shak-B*^{L41} allele was induced. PCR product amplified from *M56i* template DNA between PORF1 and PORF2 was mixed with wild type product from the *FM6* chromosome and run out on a 1.2% agarose gel alongside a control track of amplification products from the *shak-B*^{L41}*/FM6* genotype. Separation of 2 bands was clearly visible in the control but could not be seen in the *M56i* + *FM6* product mixture (data not shown).

Figure 3.9 shows the sequence of *shak-B*^{L41}*/FM6* from the PORF2 primer. At the bottom of the *shakB*^{L41}*/FM6* sequencing ladder, a single sequence is visible, corresponding to homozygous wild-type DNA present on both chromosomes in this region. Further up the sequence ladder, two sequences are seen to be superimposed. Both sequences are clearly legible, and subtraction of the expected wild-type sequence reveals that the superimposed bases correspond to the wild type sequence 17 bases ahead. Thus the observed polymorphism is a 17 base deletion present in the *shak-B*^{L41} chromosome.

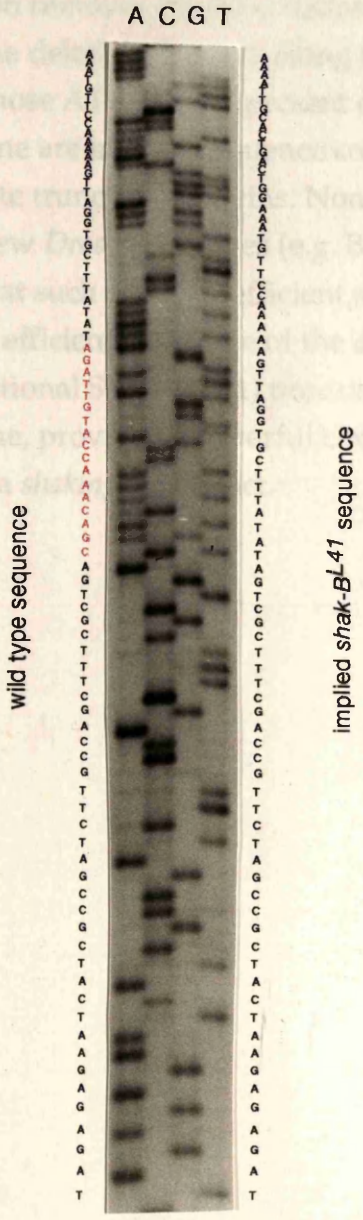


Figure 3.9. Sequencing gel showing sequence of asymmetric PCR template derived from *shak-B*^{L41}*/FM6* heterozygous females. Initial double stranded PCR amplification was between PORF1 and PORF2; the sequencing primer used was PORF1 (see figure 3.8). Reactions are loaded in order A, C, G, T. Letters on the left show the wild type sequence as derived from analysis of cloned, wild type, genomic DNA. All other genotypes sequenced (except *shak-B*^{EF535}*-see text*) contained only this wild type sequence (not shown). Subtraction of this wild type sequence from the two superimposed sequences shown yields the sequence of the *shak-B*^{L41} chromosome, as spelled out to the right of the gel photograph. Comparison of this mutant sequence with the wild type reveals a 17 base deletion in the mutant. The bases absent from the *shak-B*^{L41} chromosome are shown in red in the wild type sequence.

The exact structure of the deletion in *shak-B*^{L41} is shown in Figure 3.10. Significantly, the mutation removes the ATG codon that initiates the Shak-B(14.1) reading frame. The deletion does not bring any other upstream ATG codons into frame, and those ATG codons present downstream within the Shak-B(14.1) reading frame are in poor sequence contexts for translation, and would, in any case, initiate truncated proteins. Non-ATG initiation codons have been reported in a very few *Drosophila* genes (e.g. Bellen, *et al.*, 1992; Shugihara, *et al.*, 1990) but initiation at such sites is inefficient and it seems unlikely that such a site could operate efficiently in place of the deleted ATG codon. Thus it is very unlikely that functional Shak-B(14.1) product could be produced from the *shak-B*^{L41} chromosome, providing powerful circumstantial evidence that this protein genuinely is a *shaking-B* product.

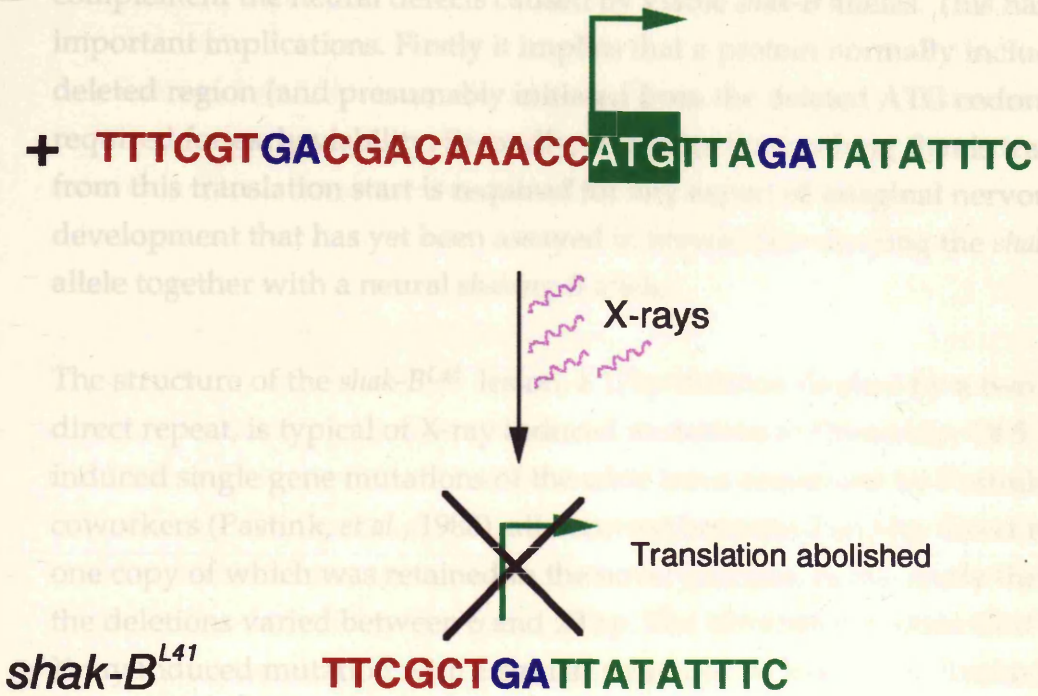


Figure 3.10. Diagram of DNA lesion underlying the *shak-B*^{L41} allele. The lesion is a 17 bp deletion flanked by a GA direct dinucleotide repeat, a structure wholly typical of X-ray induced mutations in *Drosophila* (Pastink et al., (1988); Weinzierl et al., (1987)). The deletion does not bring any potential upstream ATG codons into frame. Those ATG codons present downstream are in poor sequence contexts for translation initiation, and translation from them would, in any case, yield N terminally truncated products. Thus it is likely that no functional Shak-B(14.1) product is translated from the *shak-B*^{L41} chromosome.

While the *shak-B*^{L41} allele is lethal, it does, as discussed above (§2.2), fully complement the neural defects caused by viable *shak-B* alleles. This has two important implications. Firstly it implies that a protein normally including the deleted region (and presumably initiated from the deleted ATG codon) is required for early viability. Secondly, it implies that nothing that is translated from this translation start is required for any aspect of imaginal nervous system development that has yet been assayed in female flies carrying the *shak-B*^{L41} allele together with a neural *shaking-B* allele.

The structure of the *shak-B*^{L41} lesion, a 17bp deletion flanked by a two base pair direct repeat, is typical of X-ray induced mutations in *Drosophila*. Of 5 X-ray induced single gene mutations of the *white* locus sequenced by Pastink and coworkers (Pastink, *et al.*, 1988), all occurred between 2 or 3 bp direct repeats, one copy of which was retained in the novel junction. In this study the size of the deletions varied between 6 and 29 bp. The *Ultrabithorax* allele *Ubx*$^{6.28}$ is an X-ray induced mutation with a similar structure. 32 base pairs, flanked by a 4bp direct repeat, are deleted in this allele (Weinzierl, *et al.*, 1987). This recurring structural theme among X-ray induced alleles strongly suggests that a recombination process operates in flies to repair X-ray induced DNA damage (Pastink, *et al.*, 1988).

As mentioned above, the mutant stock thought to be *shak-B*^{EF535} was also found to carry the 17 bp deletion present in *shak-B*^{L41}. *shak-B*^{EF535} is an ethylmethane sulphonate (EMS) induced allele. EMS is a monofunctional alkylating agent which primarily induces GC to AT transitions, due to the mispairing of O^6-alkyl G with T (Snow, *et al.*, 1984). It is very unlikely that the observed 17bp deletion allele was induced by EMS. As mentioned above, the lesion is very typical of X-ray induced mutations, thus the most likely explanation is that the "*shak-B*^{EF535}" stock in fact contains *shak-B*^{L41}. In an attempt to resolve this issue, the same region was sequenced from different *shak-B*^{L41} and *shak-B*^{EF535} stocks. These were the *shak-B*^{L41}*;bw;st* and *shak-B*^{EF535}*;bw;st* stocks which were generated several years ago in the laboratory of Robert Wyman to test the escape responses in flies carrying *shak-B* alleles. The result was the same. Both stocks carried the same mutation, suggesting that the "*shak-B*^{EF535}" stock has in fact been *shak-B*^{L41} for several years, unbeknown to its tender custodians. This observation may account for the rather similar phenotypic manifestations of these two alleles (Baird, *et al.*, 1990).

3.5.5.c The search for *shak-B* mutant lesions: The 80 codon ORF of KE2(1.8)

Because the *shak-B*^{L41} mutation fully complements the viable, neural alleles of *shaking-B*, the lesions underlying these alleles must disrupt distinct genetic functions. The structure of the *shak-B*^{L41} mutation thus implies that presumptive neural proteins, disrupted by the *shak-B*^{2} and *shak-B*^{Passover} lesions cannot be initiated from the same translation start as Shak-B(14.1). This being so, it seemed at least worth investigating whether another prominent ORF of KE2(1.8) might have a neural function. This would not be immediately consistent with the genetic predictions made for the relationships between the essential and neural functions of *shaking-B*, as these functions were anticipated to have a region of overlap (§3.3), but it seemed plausible that this overlap could reside in *cis*-acting control regions driving expression of the KE2(1.8) transcript, or that alternative splicing at *shak-B* could generate transcripts in which two separate ORF regions are joined to common downstream coding sequences. For this reason, the 80 amino acid ORF region was screened for lesions which might cause the *shaking-B*^{Passover} and *shaking-B*^{2} alleles. In this experiment, PCR was performed on *shaking-B*^{Passover} and *shaking-B*^{2} DNA templates (hemizygous and homozygous templates, respectively), between primers P1 and P2 (§2.10.3; see Figure 3.6 for primer positions). This PCR amplifies a 502 bp fragment when wild type template DNA is used. Amplification and sequencing of this fragment by an asymmetric PCR strategy identical to that described above, demonstrated that *shaking-B*^{Passover} and *shaking-B*^{2} are wild-type throughout the 80 codon ORF[1].

3.5.6 *shaking-B* cDNA P1

Using KE2(1.8) as a probe, further cDNA library screening was undertaken in order to isolate more *shak-B* cDNAs (Jane A. Davies and Helena Yang, unpublished data). In order to determine the number and sizes of transcripts sharing common sequences with KE2(1.8) and to evaluate their expression levels at different stages of development, Northern analysis using a KE2(1.8)

---

1 Although the discovery of a lethal *shaking-B* lesion in this region was considered highly unlikely, the P1-P2 interval was also amplified from template DNA from flies carrying the lethal alleles *shak-B*$^{R-9-29}$, *shak-B*^{EC201}, *shak-B*^{E81}, *shak-B*^{HM437} and *shak-B*$^{17-360}$. Agarose gel analysis did not reveal any polymorphisms in these PCR products (not shown).

probe was attempted (Alan Griffin, unpublished data). Despite many careful attempts and the successful detection of numerous control RNA species, transcripts homologous to KE2(1.8) were never detected. This is a reflection of the rarity of *shaking-B* transcripts, as was later apparent both from cDNA library screening experiments (Shuqing Ji and D. E. C., unpublished data), and from *in situ* hybridisation experiments performed by Martin Todman (Crompton, *et al.*, 1995).

The rarity of *shaking-B* cDNAs in all libraries tested made the isolation of clones by conventional screening methods very arduous (Shuqing Ji, unpublished ardure). Two cDNA species, termed N52 and P1 were, however, successfully recovered by standard methods (Jane A. Davies and Helena Yang, unpublished results; Shuqing Ji, unpublished results). The N52 cDNA is discussed in Chapter 4.

The P1 cDNA was isolated from an adult cDNA library (Stratagene) (Jane A. Davies and Helena Yang, unpublished results). The cDNA is 3.9 kb in length, and detailed restriction mapping revealed the structure shown in Figure 3.11. Much of this pattern resembled the structure of the psBGC and psBGD genomic subclones (see Figure 3.6) generated for the structural analysis of KE2(1.8), and, in this region at least, P1 appeared colinear with the genome. This immediately suggested that P1 might not be a *bona fide* cDNA, but rather a genomic DNA fragment, a suspicion that was heightened by the deplorable levels of genomic DNA contamination observed in the Stratagene cDNA library from which the clone was isolated. Further detailed restriction mapping of the genomic region from which P1 was derived revealed that, at the level of resolution afforded by agarose gel separation of restriction fragments (data not shown), P1 seemed entirely colinear with the genomic DNA. One piece of evidence does, however, suggest that P1 is a real cDNA clone. At the 3' (distal) end of the cDNA, a poly(A) sequence of 46 residues was found. Sequencing of the same region in the p94.R1 genomic subclone did not reveal any series of A residues. As the poly(A) sequence observed in the cDNA is very much longer than the oligo(T) primer used during library synthesis to initiate first strand cDNA synthesis, it appears that the poly(A) segment is derived from a poly(A) tail added post-transcriptionally, implying that P1 is a genuine cDNA. P1 does not, however, show a canonical poly(A) signal (AATAAA, (Proudfoot and Whitelaw, 1988)), or any of its common variants. While this is unusual for a polyadenylated transcript, it has been suggested that inefficient polyadenylation, due to the

absence of consensus polyA signals, may be used by some loci as a means of control of gene expression (Proudfoot and Whitelaw, 1988), and it may be that the 3' end of P1 harbours a sequence so divergent from the consensus as to be unrecognisable. The 5' end of P1 does not show an extra G residue, thus P1 may or may not be full length.

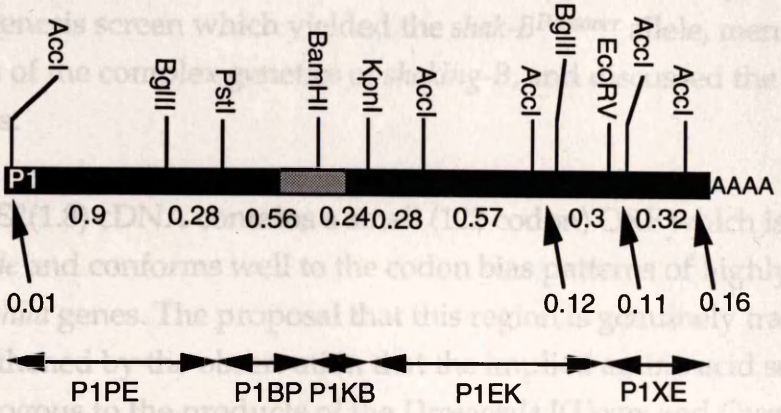


Figure 3.11: Restriction map of the P1 cDNA. Restriction sites are shown above the bar representing the P1 sequence, while fragment sizes are shown below. Blue arrows show the extents of the five subclones constructed to facilitate sequencing.

P1 was subcloned into 5 fragments, as shown in Figure 3.11 above, and these were partially sequenced using both T7 and T3 primers and custom made primers dedicated to the insert sequences (as shown in Figure 4.8). The sequences so derived are presented in the appendix to chapter 4. The full sequence of P1 has not been elucidated. P1 has, however, been sequenced almost entirely on one strand, and this analysis has demonstrated unequivocally that no large ORF exists. P1 overlaps with KE2(1.8) throughout exon B of the latter, including the coding region harbouring the *shak-B*^{L41} deletion. The first 120 codons of this reading frame are common to both cDNAs, but the sequences diverge at the point where KE2(1.8) splices from exon B to exon C. The alternative splicing of this reading frame was of great potential interest: given the position of the *shak-B*^{L41} mutation it was anticipated that alternate splicing at this junction could perhaps yield a longer protein product. However, while the reading frame of KE2(1.8) is found to stop 2 codons after the splice junction, the P1 sequence, continuing into the intron of KE2(1.8) encounters a stop after only a single codon. Thus the 122 residue potential product of KE2(1.8) is identical to the 121 residue ORF of P1 except for

its last two residues. The predicted molecular weight of the P1 product is 14.0 kDa, and it will be referred to as Shak-B(14.0).

3.6 Summary

In this chapter I have described the *Drosophila* giant fibre system and the mutagenesis screen which yielded the *shak-B*^{Passover} allele, mentioned some details of the complex genetics of *shaking-B*, and discussed the structures of two cDNAs.

The KE2(1.8) cDNA contains a small (122 codon) ORF which is highlighted by *Testcode* and conforms well to the codon bias patterns of highly biased *Drosophila* genes. The proposal that this region is genuinely translated is strengthened by the observation that the implied amino acid sequence is highly homologous to the products of the *Drosophila l(1)ogre* and *Caenorhabditis elegans unc-7* genes. Circumstantial evidence that this proposed product of KE2(1.8) is indeed a Shaking-B protein comes from the observation that both homologues are themselves involved in nervous system development, while results showing that the Shak-B(14.1) translation start is removed by a small deletion associated with the *shak-B*^{L41} allele provides more compelling evidence. This putative protein is referred to as Shak-B(14.1).

The 3.9 kb adult cDNA P1 contains a slightly smaller (121 residues) reading frame which overlaps with 120 amino acids of the Shak-B(14.1) coding sequence. Despite its large size, it contains no larger stretches of reading frame. The P1 cDNA appears not to have any introns, but does have a genuine polyA tail, and thus is accepted as a real cDNA rather than a contaminating genomic DNA fragment. The putative product of P1 is referred to as Shak-B(14.0).

The homology described above between Shak-B(14.1) and Ogre is every strong indeed, but it is striking that Ogre is much longer than its Shak-B homologue, and that all of the homology resides within the first 121 amino acids of the 362 residue Ogre protein. As briefly alluded to above, careful observation of DNA and protein level searches reveals similarities between *l(1)ogre*, *unc-7* and exon D of KE2(1.8), a region not thought to be coding in this splice form. It was soon noticed that translation of a transcript *lacking exon C* would yield a longer product with extended homologies to both the Ogre and Unc-7 proteins. Such a

splice variant was sought and found, and this and other *shak-B* transcript forms, will be presented in chapter 4.

Four

4.1 THE B10 cDNA FRAGMENT

Given the observation (see chapter 3) that conceptual translation of a *shak-B* cDNA skipping exon C of KE2(1.8) would yield a product with an extended region of homology to Ogre and Unc-7, such a clone was sought in cDNA libraries. The initial approach (Jane Davies, unpublished results) was to use a primer (§2.10.3) pointing upstream from within exon D (PCRP2) and a primer facing downstream from exon B (PCRP1) in PCR experiments on cDNA library aliquots (see Figure 4.8 for primer positions). A PCR approach was desirable due to the rarity of cDNA clones which had repeatedly been observed. An amplification product of the size anticipated for the exon C-skipping *shak-B* splice form was detected in a 12-24hr embryonic cDNA library constructed by Nick Brown in his pNB40 vector (Brown and Kafatos, 1988; data not shown).

A primer (NBT7) was designed according to the pNB40 vector sequence, facing upstream from beyond the 3' ends of the directionally cloned inserts. In PCR experiments using this primer in conjunction with PCRP1, Marian Wilkin (unpublished results) isolated and cloned a 1.2 kb fragment representing a substantial fraction of the cDNA of interest. This cDNA fragment was named B10. It was subcloned, as shown in Figure 4.1, below, and the resulting inserts were sequenced on both strands (this work). The sequence of this clone can be found in the appendix to this chapter.

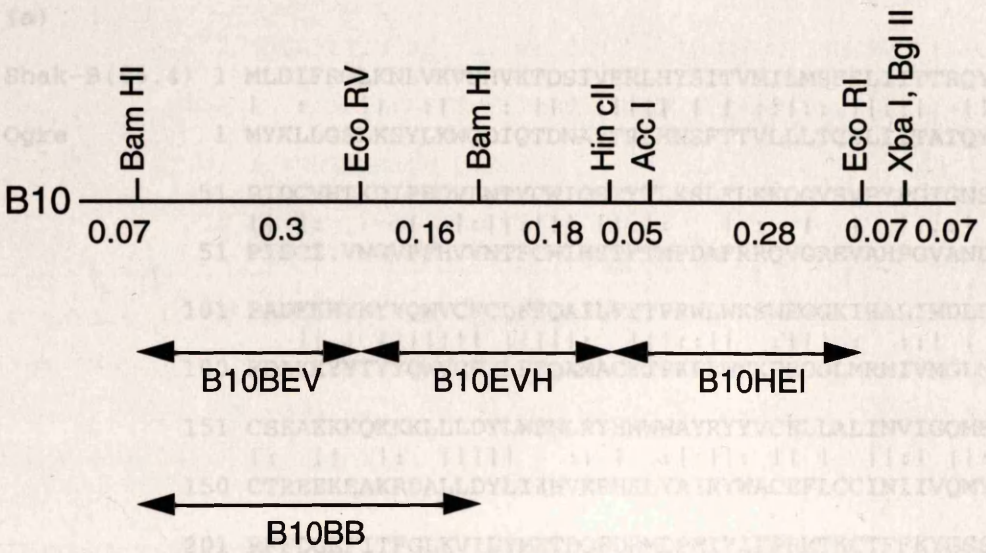


Figure 4.1: Restriction map of the B10 cDNA fragment showing regions subcloned for sequencing (arrows), and the names given to the subclones generated. Numbers indicate the sizes of restriction fragments.

For the moment we may assume exon B of KE2(1.8) is present in an identical form in the transcript from which B10 is derived, an assumption whose validity will be demonstrated later in this chapter. Given this assumption, analysis of the B10 sequence implies a protein of 372 amino acids, showing extended homologies with both Ogre and Unc-7, as shown in Figure 4.2, and having its first 120 amino acids in common with Shak-B(14.1). The predicted molecular weight of this product is 44.4 kDa, and it will therefore be referred to as Shak-B(44.4). Like the 14.1 kDa predicted product of KE2(1.8), Shak-B(44.4) is a basic protein (pI=9.2). Shak-B(44.4) and Ogre show 47% identity across the entire length of both proteins, allowing 1 gap. When chemically conservative changes are tolerated, the level of homology climbs to 65%. The *Rdf2* (§2.11.3) z score (Pearson and Lipman, 1988) for this homology is 95. Shak-B(44.4) shows 26% identity to amino acids 121-519 of Unc-7, when 12 gaps are permitted. With conservative changes, the Shak-B(44.4) - Unc-7 homology is 43%. For this homology, z = 17. Recall that a z score greater than 6 is considered highly significant.

(a)

```
Shak-B(44.4)   1 MLDIFRGLKNLVKVSHVKTDSIVFRLHYSITVMILMSFSLIITTRQYVGN  50
Ogre           1 MYKLLGSLKSYLKWQDIQTDNAVFRLHNSFTTVLLLTCSLIITATQYVGQ  50

              51 PIDCVHTKDIPEDVLNTYCWIQSTYTLKSLFLKKQGVSVPYPGIGNSDGD 100
              51 PISCI.VNGVPPHVVNTFCWIHSTFTMPDAFRRQVGREVAHPGVANDFGD  99

             101 PADKKHYKYYQWVCFCLFFQAILFYTPRWLWKSWEGGKIHALIMDLDIGI 150
             100 EDAKKYYTYYQWVCFVLFFQAMACYTPKFLWNKFEGGLMRMIVMGLNITI 149

             151 CSEAEKKQKKKLLLDYLWENLRYHNWWAYRYYVCELLALINVIGQMFLMN 200
             150 CTREEKEAKRDALLDYLIKHVKRHKLYAIRYWACEFLCCINIIVQMYLMN 199

             201 RFFDGEFITFGLKVIDYMETDQEDRMDPMIYIFPRMTKCTFFKYGSSGEV 250
             200 RFFDGEFLSYGTNIMKLSDVPQEQRVDPMVYVFPRVTKCTFHKYGPSGSL 249

             251 EKHDAICILPLNVVNEKIYIFLWFWFILLTFLTLLTLIYRVVIIFSPRMR 300
             250 QKHDSLCILPLNIVNEKTYVFIWFWFWILLVLLIGLIVFRGCIIFMPKFR 299

             301 VYLFRMRFRLVRRDAIEIIVRRSKMGDWFLLYLLGENIDTVIFRDVVQDL 350
             300 PRLLNASNRMIPMEICRSLSRKLDIGDWWLIYMLGRNLDPVIYKDVMSEF 349

             351 ANRLGHNQHHRVP 363
             350 AKQVEPSKHDRAK 362
```

(b)

```
Shak-B(44.4)   2  LDIFRGLKNLVKVSHVKTDSIVFRLHYSITVMILMSFSLIITTRQYVGN  50
Unc-7        121  MILYYLASAFRALYPRLDDDFVDKLNYYYTTTILASFALLVSAKQYVGF 169

              51 PIDC...VHTKDIPEDVLNTYCWIQSTYTLKSLFLKKQGVSVPYPGIGNS  97
             170 PIQCWVPATFTDAMEQYTENYCWVQNTY.....WVPMQEDIPR 207

              98 DGDPADKKHYKYYQWVCFCLFFQAILFYTPRWLWKSW....EGGKIHALI 143
             208 EIYSRRNRQIGYYQWVPFILAIEALLFYVPCILWRGLLYWHSGINLQGLV 257

             144 ........MDLDI...GICSEAEKKQKKKLLLDYLWENLR...YHNW... 176
             258 QMACDARLMDSEIKTRTVYTMARHMQDEVQLTNIDRQGHSRSCFSNLQLG 307

             177 ....WAYRYYVCELLALI......NVIGQMFLMNRFFDGEFITFGLKVID 216
             308 ANCGRHCGCYVTMLYIGIKVLYSANVLLQFFLLNHLLGSNDLAYGFSLLK 357

             217 YMETDQEDRMDPMIYIFPRMTKCTFFKYGSSGEVEKHDAICILPLNVVNE 266
             358 DLMHAIEWEQTGM...FPRVTLCD.FEVRVLGNIHRHTVQCVLMINMFNE 403

             267 KIYIFLWFWFILLTFLTLLTLIYRVVIIFSPRMRVYLFRMRFRLV..... 311
             404 KIFLFLWFWFLTCGIVTVCNTMYWILIMFIPSQGMSFVRKYLRVLPDHPA 453
```

```
312 RRDAIEIIVRRS.....KMGDWFLLYLLGENIDTVIFRDVVQDLANRLGH 356
    :  | :: :|:       :    |:| ::       ::  :::  | :
454 KPIADDVTLRKFTNNFLRKDGVFMLRMISTHAGELMSSELILALWQDFNN 503

357 NQHHRVPGLKGEIQDA 372
              :|   :
504 VDRSPTQFWDAEHGQG 519
```

Figure 4.2: Homology between Shak-B(44.4) and (a) the Ogre protein; (b) the Unc-7 protein. Identities are shown with lines (|), conservative changes (A,G; S,T; D,E; K,R; V,I,M,L; N,Q; Y,W,F) are indicated with colons (:).

In order to identify genomic fragments harboring exons of B10, Southern blots of Eco RI digested 952 walk λ phages were probed with a ^{32}P labelled (§2.8.1, §2.9.3) B10 probe (Marian Wilkin, unpublished). In addition to those fragments already known to contain exons of KE2(1.8), three more distal Eco RI fragments were seen to hybridise (data not shown). These fragments were 1.7, 1.7, and 2.5 kb in length. Each was subcloned (this work) and finely restriction mapped with 14 enzymes (Hin dIII, Bam HI, Hin cII, Pst I, Kpn I, Sst I, Sal I, Pvu II, Bgl II, Xho I, Xba I, Sma I, Acc I and Eco RV). The three Eco RI fragment subclones were designated psBGJ, psBGK and psBGM. The positions of B10 exons were further refined by probing Southern blots of restriction enzyme digested psBGJ, psBGK, and psBGM subclones with a ^{32}P labelled B10 probe. On the basis of these results, four smaller subclones (psBGJRS, psBGKBE, psBGKHB, and psBGKEH) were generated to facilitate sequencing, on at least one strand, of the genomic DNA containing the B10 exons. The genomic organisation of B10 is shown in Figure 4.8. The final exon of KE2(1.8), truncated by internal priming, is present in a complete form in B10. Further downstream, four novel exons are found. The last of these ends in a short (11 residue) poly(A) sequence also present in the genome at the same position. No recognisable poly(A) signal (Proudfoot and Whitelaw, 1988) is found, thus making it probable that this cDNA, like KE2(1.8), was primed from an internal A-rich sequence.

The full sequence of the B10 fragment is presented in the appendix to this chapter. The structure of the implied peptide sequence of B10 will be discussed in chapter 5.

4.2 AN INVERSE PCR SCREEN FOR *SHAK-B* cDNAS

As mentioned above, the rarity of *shak-B* cDNA clones in cDNA libraries makes library screening by conventional techniques very arduous. It is notable that, at the time of writing, seven of the eight *shak-B* cDNA splice variants isolated to date have been internally primed. This raises the possibility that *shak-B* mRNAs are inefficiently polyadenylated or that transcripts have very large, untranslated 3' ends such that priming from the poly(A) tail seldom generates cDNA clones long enough to include coding sequence. The method used to isolate B10 is useful, but careful consideration reveals some shortcomings. First, and least significant, is the problem, common to all Taq polymerase-based PCR methods, of polymerase errors. These are not a major concern in the context of the molecular analysis of *shak-B* because all PCR cDNA sequences can be carefully double checked against the sequences of corresponding genomic subclones. A greater problem is that if differential splicing is occurring in multiple exons, then amplification of 5' and 3' ends from cDNA libraries must be followed by matching different 5' transcript starts and splice patterns to differentially spliced or processed 3' fragments, in order to confidently assemble the sequences of full-length clones. A method of circumventing this problem was therefore devised.

Because it yielded the B10 sequence, the 12-24 hr embryonic cDNA library constructed by Nick Brown (Brown and Kafatos, 1988) became the library of choice for the search for a full length version of this and other *shak-B* cDNAs. This library was constructed in the 2.49 kb purpose-built plasmid vector pNB40. Because this vector is so small, inverse PCR from adjacent, divergent, gene specific primers was considered to be feasible, enabling simultaneous amplification of each flank of the insert together with the intervening vector backbone. Intramolecular ligation of the ends of such clones should yield a gene-specific library of plasmids which may be recovered by electroporation of *E. coli* and growth, on selective medium, of cells containing plasmids.

While conceptually simple, this technique demands amplification of rather large (perhaps 3-6 kb) fragments and this required careful optimisation of PCR conditions. The KE2(1.8) cDNA in pBluescript, yielding a 4.7kb product molecule, was used as a control template on which to optimise PCR reaction conditions. Initially both Taq polymerase (obtained from Promega) and Pfu™ polymerase (Stratagene) were tested. The manufacturers claim a very high processivity and fidelity of replication for Pfu polymerase, relative to Taq. In

my hands, however, Pfu gave only unsatisfactory smears, and Taq was adopted as the enzyme of choice. A PCR buffer based on tricine (rather than tris) has been recommended for the amplification of large products with Taq polymerase (Ponce and Micol, 1992). Initial trials used this buffer in parallel with the Promega 10x buffer used routinely for PCR (§2.12.1). Cycling temperatures were adjusted for different buffers according to the method described (§2.10.1). The Promega buffer gave more consistent results with the control template and was therefore used in the screening experiment itself.

4.2.1 Choice of primer pair for inverse PCR screen

The primer pair P5 and P11 (refer to Figure 3.6 for the positions of these primers) was selected for an inverse PCR cDNA screen. These primers lie within the small 80 amino acid ORF of KE2(1.8). This region was chosen because it was evident that the translation start used by the KE2(1.8) 122 amino acid ORF (and probably also by the B10 ORF) was shown by the position of the *shak-B*^{L41} lesion not to be the start of a Shak-B(neural) protein (§3.5.5.b). It seemed at least remotely possible that the upstream 80 amino acid ORF might, as part of a larger reading frame in a splice form distinct from KE2(1.8), P1 and B10, encode a neural protein.

4.2.2 Optimisation of PCR conditions

A series of controls was carried out using KE2(1.8) as a PCR template in order to optimise PCR conditions for the P5 and P11 primer pair before an attempt was made to amplify an aliquot of cDNA library. The predicted melting temperatures of P5 and P11 are 66.5°C and 62.5°C respectively (§2.10.3). Optimal cycling conditions, yielding a single, strong, 4.7 kb band from control template, were as follows:

Cycle no.	Denaturation	Annealing	Extension
1	94 °C, 3 mins	67 °C, 1 min	72 °C, 2 mins
2,3	94 °C, 1 min	66 °C, 1 min	72 °C, 2 mins
4,5	94 °C, 1 min	65 °C, 1 min	72 °C, 2 mins
6,7	94 °C, 1 min	64 °C, 1 min	72 °C, 2 mins
8,9	94 °C, 1 min	63 °C, 1 min	72 °C, 2 mins
10-25	94 °C, 1 min	62 °C, 1 min	72 °C, 2 mins

This strategy of progressively lowering the annealing temperature in the initial stages of a PCR is known as touchdown PCR. The rationale is to give specific reaction products a head start over any spurious, contaminating bands by starting the reaction with an annealing temperature so high that little or no annealing can occur, and reducing the annealing temperature gradually, over a number of cycles through the effective annealing temperature of the primer pair.

Figure 4.3 shows the result of amplification of control plasmid DNA and of an aliquot of 12-24 hr embryonic cDNA library using the above cycling conditions, and standard reagent concentrations (§2.12.1).

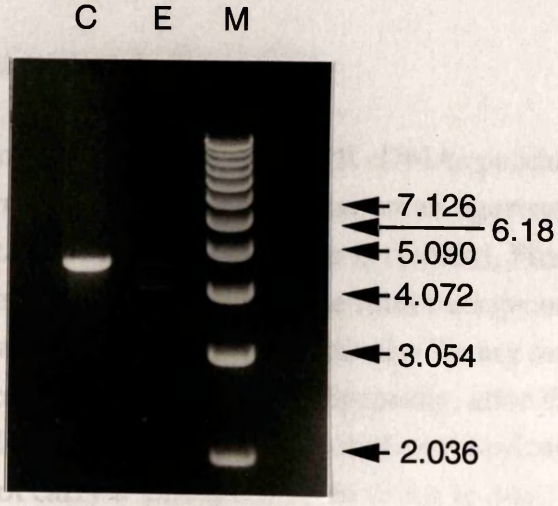


Figure 4.3. Photograph of an ethidium bromide-stained agarose gel showing inverse PCR amplification products. Lane C: control sample derived from amplification of the KE2(1.8) cDNA in pBluescript, between primers P5 and P11, yielding a single 4.7 kb product band. Lane E: experimental lane, showing amplification products derived from PCR between P5 and P11, using Nick Brown's 12-24 hr embryonic cDNA library as a template. Three product bands are clearly visible, in the 4-5 kb range. Lane M: 1kb DNA ladder (BRL). The sizes of relevant bands are shown to the right of the photograph.

4.2.3 Cloning of inverse PCR products

The initial step in cloning the inverse PCR cDNAs produced by the above PCR reaction was to run out the whole reaction on an agarose gel and purify the bands of interest. The reason for this step is twofold. Firstly, running out the reaction products in this way reduces the final background of nonspecific ampicillin-resistant plasmids derived from the library template rather than from the specifically amplified cDNAs. Secondly, after the PCR reaction, the 5' ends of the amplification products are not phosphorylated (as the primers themselves do not carry 5' phosphates). In order to enable intramolecular recircularisation, these ends must be phosphorylated using T4 polynucleotide kinase. The agarose gel purification step removes unincorporated PCR primers which would otherwise compete for kinase activity in this step.

After gel purification of the amplified bands (§2.5.1) the cDNA was precipitated using 1µl glycogen solution (20 µg), 0.1 volumes 3M NaAc pH 5.3 (N.B. NH_4^+ salts must not be used here, as they strongly inhibit T4 kinase, (Sambrook, *et al.*, 1989)), and 2 volumes of ethanol. Precipitation was carried out overnight at -20° C, and the pellet, made visible by the glycogen carrier, was recovered as described in chapter 2 (§2.5.5). Taq polymerase tends to yield products with single base 3' overhangs, making it difficult to recircularise these molecules. A variety of techniques has been proposed to circumvent this problem (e.g. Kovalic, *et al.*, 1991). Perhaps the simplest option available is to blunt-end the molecules with the 3'-5' exonuclease activity of Klenow polymerase (Sambrook, *et al.*, 1989). The blunt ending and phosphorylation steps were done simultaneously in a reaction cocktail containing the following components: resuspended PCR product (10µl), Promega 10x ligase buffer (2µl), dNTP stock (2mM each dNTP, 2µl), T4 polynucleotide kinase (1µl = 8U), Klenow polymerase (1µl = 5U), volume made up to 20µl with dH_20. This mixture was incubated for 2 hrs at 37°C, whereupon 1µl (=3 Weiss units) T4 DNA ligase was added. The incubation was then continued overnight at 16°C to allow recircularisation of molecules. The success of the ligation reaction was checked using a control ligation of PCR product from the amplification of the KE2(1.8) plasmid. The products of the control reaction were run on a 0.8% agarose gel alongside some unligated control PCR product, and the ligation was seen to have been successful (data not shown).

The products of the experimental ligation were then precipitated (§2.5.5) and resuspended in 10µl dH_2O. The purpose of this precipitation is to maximise the time constant (by minimising the ionic concentration of the cell suspension) in the electroporation reaction. The resuspended DNA was then used to electroporate XL1-Blue cells (§2.6.1.c, §2.6.2.b). Between 300 and 400 Ap^r colonies were recovered. These were tested by hybridising colony lift filters (§2.9.2, §2.9.3) with ^{32}P labelled KE2(1.8) and B10 probes (Jane Davies, unpublished results). Approximately half of the clones were found to hybridise strongly to one or both probes. Glycerol stocks were made of all positive clones and 81 positive clones were initially selected for analysis.

As would be anticipated, this PCR selection technique yields large numbers of identical clones. When compared with conventional library screening techniques, the burden of work in the inital isolation of rare clones is reduced, while the effort required in the characterisation of positives is increased somewhat. Much of the initial characterisation of positive clones was undertaken by Jane Davies, Shuqing Ji, and Mary Gardiner (unpublished results). After my initial characterisation of the SIPC8 clone (see below), their work identified a variety of clones with restriction patterns distinct from those already encountered in *shak-B* cDNAs. Regions showing novel restriction patterns were sequenced to identify novel splice commitments (this work). This analysis demonstrated several new *shak-B* transcript forms, which will be presented below. It should be stressed that minor alterations in splicing that did not significantly alter the restriction pattern of cDNAs could have gone undetected. This shortcoming can only be remedied by sequencing each of these cDNA clones in their entirety: a laborious task unlikely to repay the investment of effort. Inverse PCR cDNA clones were grouped according to their restriction patterns and a representative of each group will be discussed. The detailed structures of all of the different cDNA clones presented in this chapter, together with those discussed in chapter 3 are shown in Figure 4.8.

4.3 INVERSE PCR cDNAs ENCODING SHAK-B(LETHAL) PROTEINS

The structures of all *shak-B* cDNAs isolated to date are summarised in Figure 4.8.

4.3.1 The SIPC8 cDNA

Two of the classes of clones isolated from the inverse PCR screen contained within them the B10 cDNA fragment. The first of these was designated SIPC8 (*shaking-B* Inverse PCR cDNA 8). The restriction map of SIPC8 was determined using 15 restriction enzymes (Eco RI, Bam HI, Eco RV, Hin cII, Hin dIII, Acc I, Pvu II, Pst I, Sal I, Sma I, Sst I, Xba I, Xho I, Bgl II, and Kpn I), and this mapping revealed a pattern consistent with a cDNA containing all of the B10 fragment plus the upstream exons of KE2(1.8). SIPC8 was subsequently sequenced on one strand throughout its coding region. The 5' and 3' untranslated extremities were also sequenced. The clone was found to contain the Shak-B(44.4) 372 amino acid open reading frame anticipated from the B10 sequence. The 5' end of the cDNA was identical to that of KE2(1.8) except that the first base of the SIPC8 clone is one base distal to that of KE2(1.8). SIPC8, like KE2(1.8) has an extra G residue as its first base, thought to represent a reverse-transcribed 7-methyl guanosine cap. Thus SIPC8 appears to contain the full 5' extent of the transcript from which it was derived. At its 3' end, SIPC8 appears to have been internally primed at the same site as B10.

4.3.2 The SIPC737 cDNA

In the KE2(1.8) and SIPC8 cDNA clones, the splice junction between exons A and B creates a Hin cII site. This site was found to be absent from the class of cDNA represented by the SIPC737 clone (Shuqing Ji, unpublished results) suggesting that the SIPC737 clone might have an altered splice commitment in this region. Sequence analysis of SIPC737 (this work) revealed that this indeed was the case, though the difference is only very slight. The two 5' splice donor sites at the end of exon A are identical in SIPC8 and SIPC737. The 3' splice acceptors at the start of exon B are, however, different, that of 737 being three bases upstream of (proximal to) the splice acceptor of SIPC8. This difference between the splice sites of the two clones accounts for the presence or absence of the Hin cII site (see Figure 4.8). From its high resolution restriction map, the SIPC737 clone appears to be otherwise identical to SIPC8.

4.3.3 The SIPC726 cDNA

The restriction pattern of the SIPC726 cDNA shows it to be truncated at its 5' and 3' ends, relative to SIPC8. Sequence analysis revealed that not only is the clone truncated, but it demonstrates novel splice commitments at its 5' and 3'

ends. SIPC726 is found to contain all of exon E, but not to make the exon E to exon F splice, and to continue for 87 bases before ending in an A-rich sequence. This sequence is present in the corresponding genomic region, is not preceded by a recognisable poly(A) signal, and thus SIPC726 is likely to have been internally primed. All of exon E is coding in the B10, SIPC8, and SIPC737 clones. In SIPC726, the extended exon E (designated exon E') encodes only two further residues before encountering a stop codon. This makes the implied SIPC726 peptide 196 residues long. The molecular weight of this predicted product is 23.2 kDa; its pI is 9.1.

At its 5' end, SIPC726 starts in the genomic region corresponding to the A to B intron of KE2(1.8), SIPC8, and SIPC737. 726 starts 149 bases proximal to the start of exon B of KE2(1.8) and does not start with an untemplated G residue, thus may or may not include the full 5' extent of the transcript from which it was derived. While this difference in transcription start may be interesting in the context of the regulation of expression of the transcript represented by SIPC726, the N terminus of the implied polypeptide product is identical to that of KE2(1.8), SIPC8, and SIPC737.

At the positions of the splice acceptor site of exon B and the splice donor site of exon E, SIPC726 'fails' to splice. SIPC726 does not have a poly(A) tail, and it is at least possible that it is derived from an incompletely processed transcript. It may, alternatively, be a *bona fide*, fully processed alternative *shak-B* splice form, and, having stated this caveat, I will continue to treat it as such.

4.4 THE SEARCH FOR LESIONS UNDERLYING *SHAK-B* ALLELES: PART II

4.4.1 The search for *shak-B* mutant lesions: The 372 amino acid ORF of SIPC8

The translation start and first 120 amino acids of the reading frame of KE2(1.8) are common to the 372 amino acid reading frame of SIPC8. As described in chapter 3, this coding region was screened for *shak-B* mutant lesions and the *shak-B*^{L41} mutation, (a lethal mutation which fully complements the neural *shak-B* alleles) was found to be associated with a 17bp deletion which removes the translation start site. This implies that the 372 amino acid product of SCIP8, like

the 122 amino acid product of KE2(1.8) may be a gene product esssential for viability, but is not required for the development of the adult giant fibre system. Genomic DNA from the remaining lethal alleles of *shak-B* (*shak-B*$^{R-9-29}$, *shak-B*^{EC201}, *shak-B*^{E81}, *shak-B*$^{17-360}$ and *shak-B*^{HM437}), was therefore screened for mutant lesions within exons D, E, F, G and H. This screening was done by an asymmetric PCR technique identical to that used to find the *shak-B*^{L41} lesion. The primers used in this analysis and the results obtained are summarised in table 4.1. Note that only the PCR primers used are shown in the table; additional internal primers were also used to facilitate the sequencing of asymmetric PCR DNAs.

Exon	Primer pair used in PCR	Alleles showing polymorphisms	Mutations
D	PF1 + PF3	None detected	None detected
E	PJ1 + PJ2	None detected[1]	None detected[1]
F	PK1 + PK2	None detected	None detected
G	PK1 + PK2	*shak-B*^{EC201}, *shak-B*$^{R-9-29}$	*shak-B*^{EC201}, *shak-B*$^{R-9-29}$
H	PK6 + PM1	*shak-B*^{EC201}, *shak-B*$^{17-360}$, *shak-B*^{HM437}	none detected

Table 4.1: Results of the search for mutant lesions in genomic regions equivalent to exons D to H of SIPC8, using template DNAs from five lethal *shak-B* alleles.

[1] Sequence results from this region are not of sufficient quality to rule out the possibility of a mutation being present.

4.4.2 Polymorphisms in Exon G

shak-B^{R-9-29} and *shak-B*^{EC201} both exhibit an identical polymorphism in the genomic region equivalent to exon G of SIPC8. Both alleles were induced by EMS, a monofunctional alkylating agent which primarily induces GC to AT transitions, due to the mispairing of O^6-alkyl G with T (Snow, *et al.*, 1984). Indeed, the observed polymorphism is a G»A transition. The polymorphism occurs at the base equivalent to position 1656 of the SIPC8 sequence and changes the TGG codon of W_{273} to a TGA stop codon. This change generates a

truncated SIPC8 reading frame, lacking the C terminal 100 amino acids. It therefore seems likely that this polymorphism genuinely represents a mutant lesion of *shaking-B*. It should be noted that, at the time of writing, the sequences of the *R-9-28* and *EC235* control chromosomes have yet to be determined in this region. While this is an experimental priority, the expected functional consequences of the mutation observed in *shak-B*^{EC201} and *shak-B*$^{R-9-29}$ and the absence of this polymorphism in the seven other chromosomes sequenced (*shak-B*^{2}, *shak-B*^{HM437}, *shak-B*$^{17-360}$, *shak-B*^{E81}, *FM6*, *FM7*, and *Binsn*) make it seem extremely unlikely that this mutation will be present in the *shak-B*$^{+}$ control chromosomes.

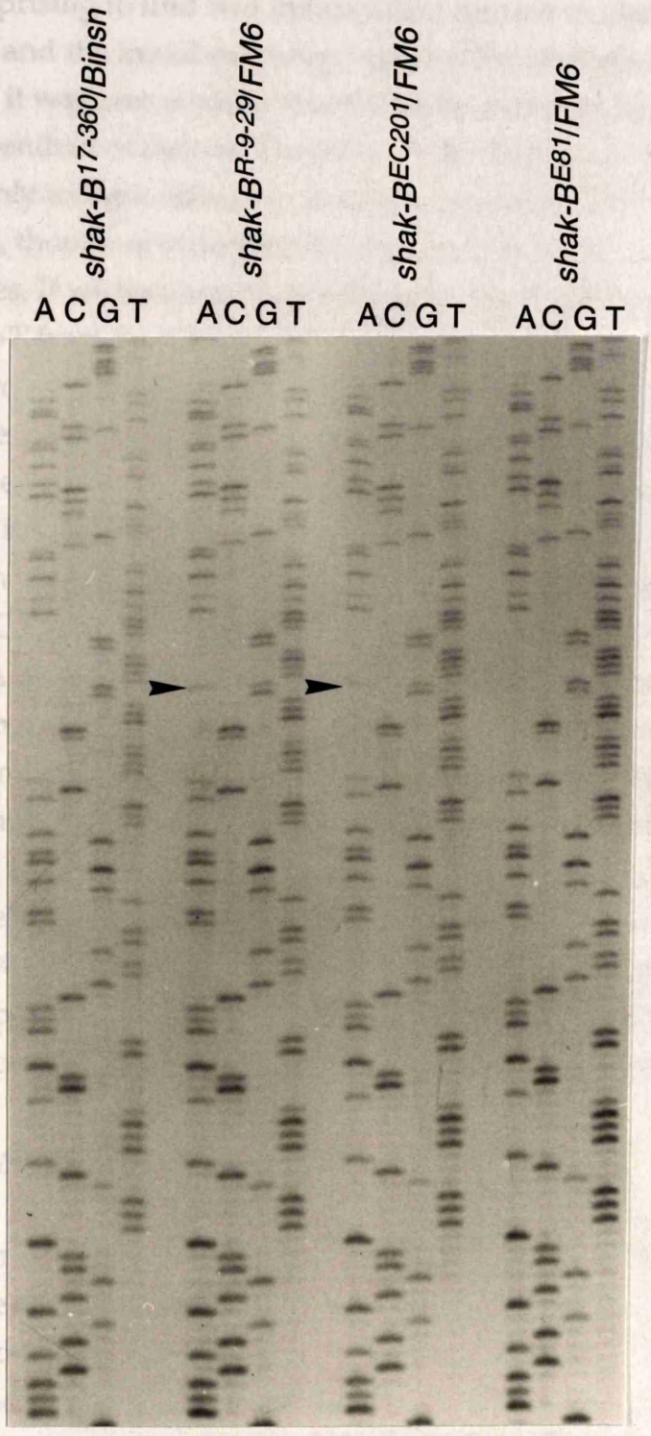


Figure 4.4. Sequencing gel photograph showing sequences of asymmetric PCR template DNAs in exon G from four different genotypes. PK4 was used as the sequencing primer. A substitution polymorphism is seen in the *shak-B^{R-9-29}/FM6* and *shak-B^{EC201}/FM6* genotypes, at the position marked with an arrow. Sequencing of cloned wild-type genomic DNA reveals a G in this position, as seen in all other genotypes sequenced (*shak-B^{2}*, *shak-B^{HM437}/FM6*, *shak-B$^{17-360}$/Binsn*, *shak-B^{E81}/FM6*). In the *shak-B^{R-9-29}/FM6* and *shak-B^{EC201}/FM6* genotypes, an A/G polymorphism is found. The A is absent from the *shak-B^{E81}/FM6* sequence ladder (and from all other genotypes tested) and thus is not derived from the *FM6* balancer chromosome. The A substitution is anticipated to create a premature stop codon (TGA), yielding a truncated (and presumably nonfunctional) protein in the mutants.

It was surprising to find two independent mutant stocks harbouring the same mutation, and the initial suspicion was that the stocks had been mixed up. However, it was also possible that the same mutation had been generated on two independent occasions. The phenotypic analysis of Baird and coworkers showed only a slight difference between the *shak-B*$^{R\text{-}9\text{-}29}$ and *shak-B*^{EC201} mutations, though nonidentical lesions could certainly confer similar phenotypes. If we temporarily pretend that the only lesion that EMS can induce is G»A (C»T from mutations on the opposite strand), and that only nonsense codon introduction will be sufficiently deleterious to cause a lethal allele, then we may derive an educated underestimate of the probability of two identical lethal alleles being independently isolated, by considering how many sequences in the reading frame region of interest may be mutated to generate stop codons by single G»A or C»T changes. Only four codons fulfill this criterion (CAA, CGA, TGG, and CAG). Of these codons, TGG is mutable by single G»A transitions both to TGA and to TAG, and must therefore be counted twice. In the region of the SIPC8 reading frame thought to be required both for essential and neural *shak-B* functions (see below), these four codons together occur twenty times. Ten are TGG codons, bringing the score of potential mutability to thirty. Thus the chances of two independently created, EMS induced, lethal, neural *shaking-B* lesions being identical in their sequence is less than[1] 3.33%. This is not very probable, but neither is it astronomically unlikely, and I will present evidence below to prove that it is, indeed the case, and that this identical transition underpins both *shak-B*^{EC201} and *shak-B*$^{R\text{-}9\text{-}29}$ alleles.

4.4.3 Polymorphisms in and around exon H

Three distinct polymorphisms are observed in the exon H region, though each lies outside the reading frame of exon H. None of these polymorphisms disrupts existing splice sites, and neither (from observation of their sequences), is any of the changes likely to enhance cryptic splice sites. These polymorphisms are thus not thought to confer mutant phenotypes. Indeed the polymorphism associated with the *shak-B*^{EC201} allele (Figure 4.7) just proximal to exon H is also found in the control *shak-B*$^{+}$ chromosome *EC235*. The detailed structures of these polymorphisms are shown in figures 4.5, 4.6 and 4.7. Their importance in the molecular analysis of *shaking-B* is twofold. Firstly the pattern of these polymorphisms proves that the *shak-B*^{EC201} and *shak-B*$^{R\text{-}9\text{-}29}$

---

1 This estimate must be an underestimate because other EMS induced lesions (e.g. deletions) are known to occur, albeit at lower frequencies.

chromosomes analysed are different, thus the existence of the same exon G mutation in both is not due to stocks having been mixed up (see below). Secondly, these polymorphisms are highly relevant to the search for the *shak-B*$^{17\text{-}360}$ and *shak-B*^{HM437} lesions (Figure 4.5). Both of these alleles were created with mutagens which tend to introduce deletions (X-rays and HMS respectively). The existence of heterozygous polymorphisms in the *shak-B*$^{17\text{-}360}$ */FM6* and *shak-B*^{HM437} */FM6* DNAs demonstrates that the region between the PK6 and PM1 primers is present in these mutant chromosomes[1].

1 Conflicting data regarding the *shak-B*^{HM437} lesion have been presented by Krishnan *et al.*, (1995), as discussed in §6.1.1.

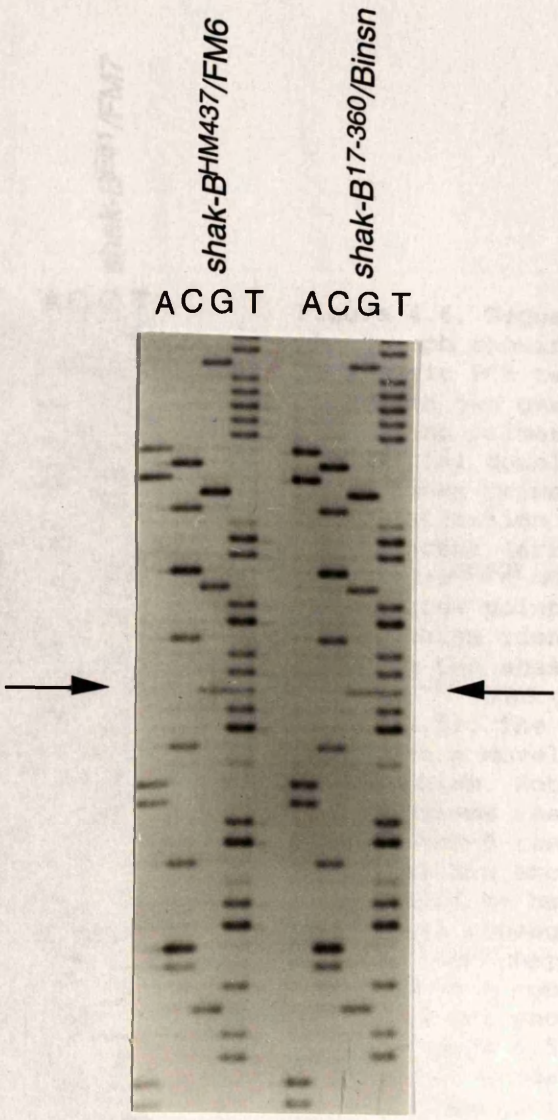


Figure 4.5. Sequencing gel photograph showing sequences of asymmetric PCR template DNAs in exon H in two genotypes. The sequencing primer used was PM1. In both sequences a G/T substitution polymorphism is apparent (arrows). The initial double stranded PCR was between PK6 and PM1 primers, thus the presence of these polymorphisms demonstrates that the exon M region is present in the *shak-B*^{HM437} and *shak-B*$^{17-360}$ alleles. The substitution itself is downstream of the Shak-B reading frame region within exon H, and is not anticipated to have any functional consequences. The same polymorphism is also present in *shak-B*^{EC201}*/FM6* (see figure 4.6), but, significantly, is absent from *shak-B*$^{R-9-29}$*/FM6* (not shown).

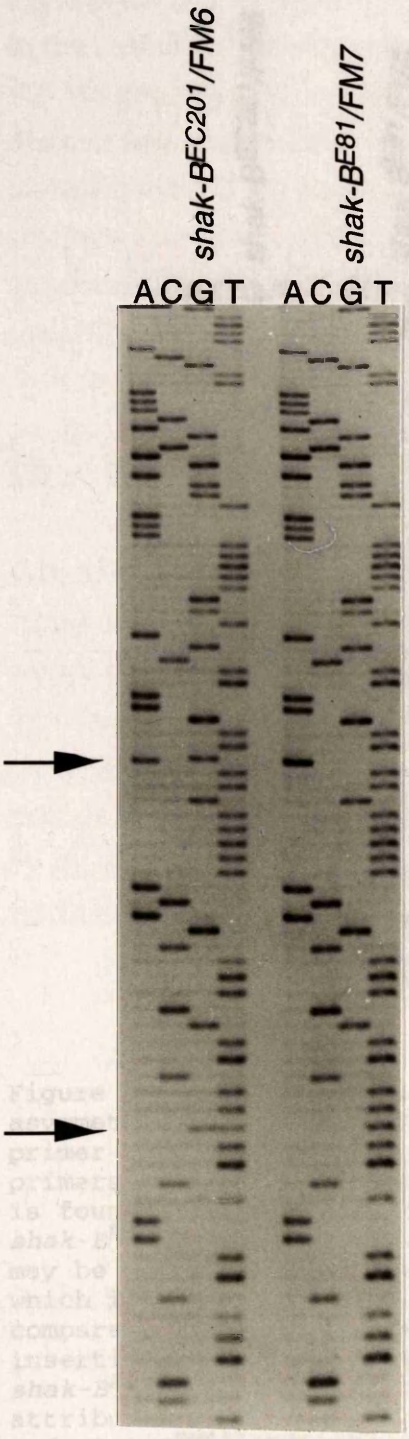


Figure 4.6. Sequencing gel photograph showing sequences of asymmetric PCR template DNAs in exon H in two genotypes. The sequencing primer used was PM1. The initial double stranded PCR was between primers PK6 and PM1. Two substitution polymorphisms are apparent (arrows), both in the *shak-B^{EC201}/FM6* genotype. The lower arrow points to a G/T polymorphism identical to that found in the *shak-B^{HM437}/FM6* and *shak-B$^{17-360}$/FM6* genotypes (see figure 4.5). The upper arrow points to a novel A/G polymorphism. Both of these polymorphisms reside downstream of the Shak-B reading frame region within exon H, and are not anticipated to have any functional consequences. The *shak-B^{E81}/FM7* sequencing ladder is included as a control for the sequencng gel photographs here and in figure 4.5. The absence of polymorphic bands in this sequence demonstrates that the extra bands found in other genotypes are caused by *bona fide* polymorphisms, rather than by compression artefacts. Neither of the polymorphisms shown here in *shak-B^{EC201}/FM6* is present in *shak-B^{R-9-29}/FM6* (not shown).

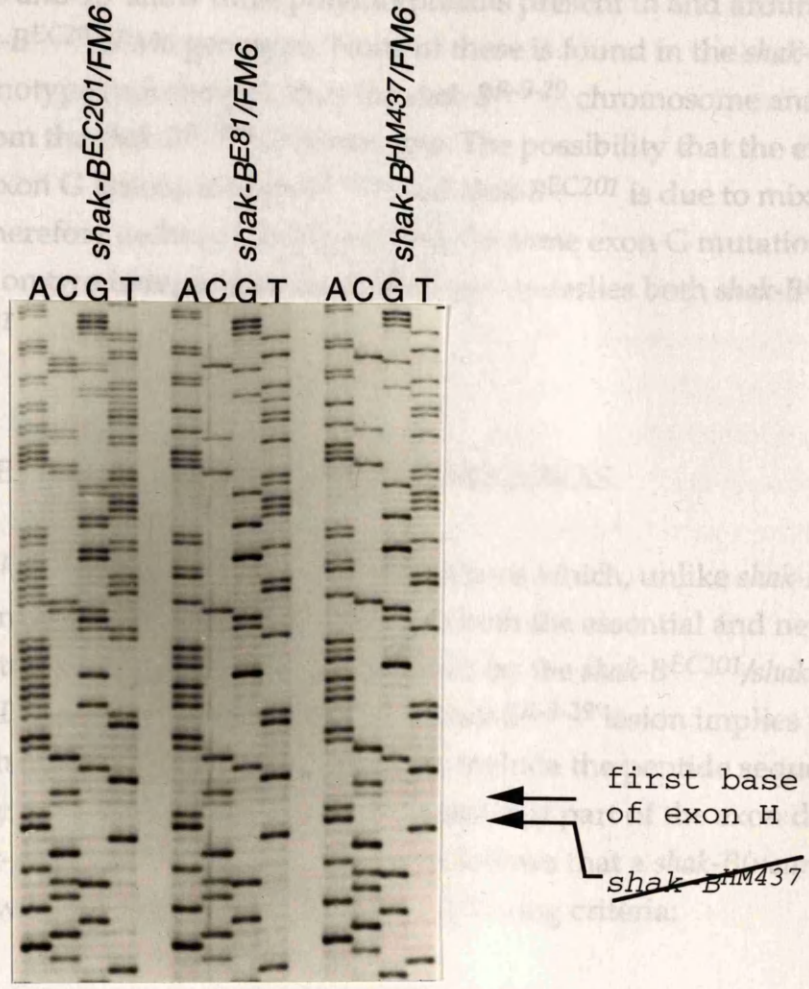


Figure 4.7. Sequencing gel photograph showing sequences of asymmetric PCR template DNAs in three genotypes. The sequencing primer used was PK8; the initial double stranded PCR was between primers PK6 and PM1. A complex deletion/substitution polymorphism is found in the *shak-B^{EC201}/FM6* but not in the *shak-B^{E81}/FM6* or *shak-B^{HM437}/FM6* genotypes. The sequence of the mutant chromosome may be inferred by subtraction of the wild type sequence upon which it is superimposed. Such analysis demonstrates that, compared to the cloned, wild-type DNA sequenced, a one base insertion and four separate base substitutions are present in the *shak-B^{EC201}* chromosome. (None of these polymorphisms is attributable to the *FM6* balancer chromosome, as the *shak-B^{E81}/FM6* and *shak-B^{HM437}/FM6* sequences demonstrate that the FM6 sequence is wild-type in this region.) The sequences shown are of the non-coding strand in the vicinity of the splice acceptor site at the start of exon H. The polymorphisms all lie upstream of exon H. The C complementary to the first G of exon H is marked with an arrow; all bases above this on the sequencing ladder are derived from the G-H intron. The proximity of the observed polymorphisms to the exon H splice acceptor site tempted the speculation that these changes might impair splicing and thus cause the *shak-B^{EC201}* mutation. Scrutiny of the mutant sequence does not, however, suggest any reason why the splice acceptor should not function normally. Furthermore, an identical polymorphism is also present in the *shak-B^{EC235}/FM6* control genotype, and thus cannot be the cause of a *shak-B* mutation. This polymorphism is not observed in *shak-B^{R-9-29}/FM6* (not shown). The *shak-B^{HM437}* lesion is not detected here (§6.1.1).

Figures 4.6 and 4.7 show three polymorphisms present in and around exon H in the *shak-B*^{EC201}*/FM6* genotype. None of these is found in the *shak-B*$^{R-9-29}$*/FM6* genotype (not shown), thus the *shak-B*$^{R-9-29}$ chromosome analysed is distinct from the *shak-B*^{EC201} chromosome. The possibility that the existence of identical exon G lesions in *shak-B*$^{R-9-29}$ and *shak-B*^{EC201} is due to mixing up of stocks is therefore excluded, implying that the same exon G mutation has been generated on two independent occasions, and underlies both *shak-B*$^{R-9-29}$ and *shak-B*^{EC201} alleles.

4.5 THE SEARCH FOR *SHAK-B(NEURAL)* cDNAS

shak-B^{EC201} and *shak-B*$^{R-9-29}$ are lethal mutations which, unlike *shak-B*^{L41}, do not complement the neural *shak-B* alleles. Thus both the essential and neural components of *shak-B* appear to be disrupted by the *shak-B*^{EC201}*/shak-B*$^{R-9-29}$ mutation. The position of the *shak-B*^{EC201}*/shak-B*$^{R-9-29}$ lesion implies that both Shak-B(lethal) and Shak-B(neural) proteins include the peptide sequences encoded by exon G (see Figure 4.8), or at least that part of the exon disrupted by the *shak-B*^{EC201}*/shak-B*$^{R-9-29}$ mutation. It follows that a *shak-B(neural)* transcript would be expected to fulfil the following criteria:

1. To include, in its coding region, some, or all of exon G.
2. To have a translation start distinct from that removed by the *shak-B*^{L41} mutation.
3. To be disrupted, at the transcriptional or translational level by neural mutations of *shaking-B*.

Such cDNA species were therefore sought.

4.5.1 The N52 cDNA

N52 was isolated by Shuqing Ji, from the Nick Brown 12-24 hr embryonic cDNA library, by conventional colony hybridisation screening. N52 was also subcloned and sequenced by Shuqing Ji. Observation of this sequence showed that the cDNA contained exons identical to D, E, F, G, and H of SIPC8, though some of its sequences 5' to this region were novel. Exon B, containing the transcription start of the 372 amino acid ORF of SIPC8, does not occur in N52. 5' to exon D, N52 does not contain any substantial regions of reading frame, and

at the start of exon D, the reading frame used in SIPC8 is not open in N52. A potential translation start does, however, exist within exon D, and the T_{Dros} score (§3.5.3.b.iii) of this start is 1.27, a relatively high value. Translation from this start would yield a 229 amino acid protein whose predicted molecular weight is 27.7. Its pI is 8.6.

When the 5' regions of N52 are compared with genomic sequences derived for analysis of KE2(1.8), it is evident that all of the psBGB genomic subclone is contained within an exon of N52. This exon continues distally beyond the end of psBGB. The adjacent 1.5 kb Eco RI fragment was therefore subcloned from λ9405, and sequenced from its proximal end (this work). This revealed that all of N52 5' to exon D was derived from a single, extended version of the exon A observed in other cDNA clones. This extended exon is designated A' (Figure 4.8). The 5' end of N52 was found to be distinct from that of KE2(1.8), SIPC8 and SIPC737. It occurs 86 bases distal to the transcription start of the SIPC8 cDNA clone, and the absence of an extra G residue at the 5' end of N52 suggests that this may be due to incomplete first strand cDNA synthesis rather than being a *bona fide* transcription start.

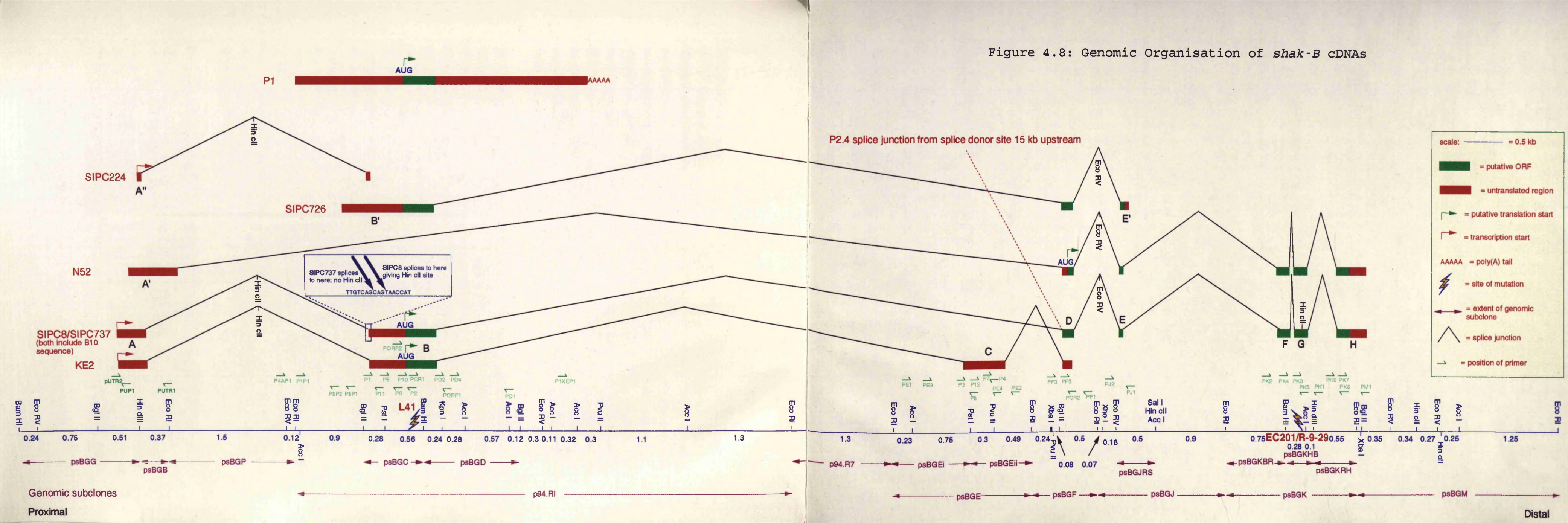
4.5.2 The SIPC224 cDNA

The SIPC224 cDNA was another of the fruits of the inverse PCR screen of the Nick Brown 12-24 hr Library. Sequence information from the 5' and 3' ends of this clone was derived by Martin Todman and I gratefully acknowledge this contribution. Restriction mapping and sequence analysis of 224 demonstrates that it contains exons G and H, but upstream of these exons its structure is quite distinct from the other cDNA clones presented. It contains a very small exon A (designated A''), starting at the base equivalent to position 306 of KE2(1.8). An extra G residue, not encoded by the genome is appended at this point suggesting that this may represent another *bona fide* transcription start. This same transcription start had been inferred by direct sequencing of PCR fragments derived from the same library in an independent experiment (Shuqing Ji, unpublished results). Sadly this evidence of a possible novel transcription start is the only useful information that may be derived from 224, as it has a large internal deletion which is likely to be a cloning artefact rather than a *bona fide* intron. Two major lines of evidence support this proposal. Firstly the sequences flanking the deleted region do not conform to the GT to AG rule (Mount, 1982). Secondly, the deletion starts immediately following the

last base of the P11 primer sequence, while the P5 primer site is deleted. Inspection of the sequences flanking this deletion do not provide a simple explanation as to a mechanism (e.g. homology to one of the PCR primers at the 3' flank of the deletion) by which this deletion may have arisen.

Figure 4.8 shows a summary diagram of all of the *shaking-B* cDNAs so far discussed.

Figure 4.8: Genomic Organisation of *shak-B* cDNAs



N52 contains the region of reading frame disrupted by the *shak-B*$^{R\text{-}9\text{-}29}$/*shak-B*^{EC201} lesion (fulfilling criterion 1, above), yet the implied translation start is not that removed by the *shak-B*^{L41} lesion (fulfilling criterion 2). N52 is therefore a candidate neural transcript of *shaking-B*. The third criterion required for a neural transcript, namely that the transcript is disrupted by lesions causing *shak-B(neural)* mutations was not satisfied by N52. Since all of the implied N52 coding sequence is also contained within the essential SIPC8 reading frame, it did not seem likely that a viable *shak-B(neural)* mutation would disrupt the N52 reading frame. There remained, however, an outside chance that *shak-B(neural)* mutations disrupted the N52 transcript form e.g. by abolishing the translation start, or by altering an N52-specific splice junction.

In a perfect world, the question of whether N52 represents a genuine *shak-B(neural)* transcript might perhaps be answered by transformation experiments to attempt rescue of the *shak-B*^{2} neural null allele. However, even if N52 does encode a Shak-B(neural) protein, a multitude of arguments why such an experiment would not work presents itself. The most compelling of these relates to genetic evidence of neural *shak-B* phenotypes being conferred by deficiencies with breakpoints tens or hundreds of kilobases upstream of the most proximal *shak-B* transcription start yet identified. This evidence will be reviewed below, but, in brief, it implies that *cis* acting control regions and/or regions of *shak-B* transcript which are yet to be isolated are required for the neural function of *shak-B*. In the absence of such upstream sequences it seems unrealistic to expect rescue of *shak-B(neural)* phenotypes by transformation with an N52 construct, even if the N52 cDNA represents a real *shak-B(neural)* transcript. Tentative identification of *shaking-B* neural transcripts therefore demanded the localisation of the lesions underlying the neural alleles *shak-B*^{2} and *shak-B*^{Passover}.

Workers in the laboratory of Robert Wyman[1] had been making parallel but independent efforts to clone and characterise *shaking-B*. Their approach had been different to that described here. They succeeded in isolating a P element allele of *shaking-B* which conferred a neural phenotype, and continued to clone the locus by transposon tagging. Their molecular analysis (Krishnan, *et al.*, 1993) described a single pupal cDNA in whose open reading frame the lesions

---

1 Despite the general consensus in favour of the name *shaking-B* (§3.3) this group continue to refer to the locus as *Passover*.

underlying the *shak-B*^{2} and *shak-B*^{Passover} alleles were shown to lie, thus fulfilling the third and most exacting criterion for *shak-B(neural)* transcript status. This reading frame, as anticipated, contained the region disrupted by the *shak-B*^{EC201}/*shak-B*$^{R-9-29}$ lesion, yet had a distinct N terminus from that affected by the *shak-B*^{L41} mutation. The cDNA (named P2.4) does not resemble N52, though it contains exons identical to exons D, E, F, and G of SIPC8. The P2.4 cDNA once again appears to have been internally primed at its 3' end, though at a site 1125 bases distal to the end of SIPC8, resulting in an extended 3' untranslated region within exon H. Upstream of exon D, however, P2.4 contains 3 exons distinct from those in the cDNAs already described. The lesions underlying the *shak-B*^{2} and *shak-B*^{Passover} alleles were both shown to lie in the coding region of the second of these exons.

The relationship between the proteins implied by P2.4 and SIPC8 is fascinating. While the anticipated product of SIPC8 is 372 amino acids in length, that of P2.4 is 361 amino acids. Its predicted molecular weight is 42.9 kDa. The 252 amino acid C terminal regions of the two proteins, derived from the same exons are, unsurprisingly, identical, except for one small polymorphism[1]. Remarkably, the two N termini, respectively 120 and 109 residues in length and derived from entirely different exons, are 69% identical. This organisation strongly suggests that these two protein forms have arisen by duplication and divergence of the N terminus of a single ancestral sequence. A multiple sequence alignment showing the homologies among the SIPC8 product, (Shak-B(44.4)), the P2.4 product (Shak-B(42.9)), and the Ogre and Unc-7 proteins is shown in Figure 4.9.

---

1 At residue 232 of Shak-B(44.4) sequence comparison to Shak-B(42.9) shows nonidentical residues. This dissimilarity is due to a DNA level A/G polymorphism at the position equivalent to base 1534 of the SIPC8 sequence, creating an ATG (methionine) codon in some fly strains and an ATA (isoleucine) codon in others.

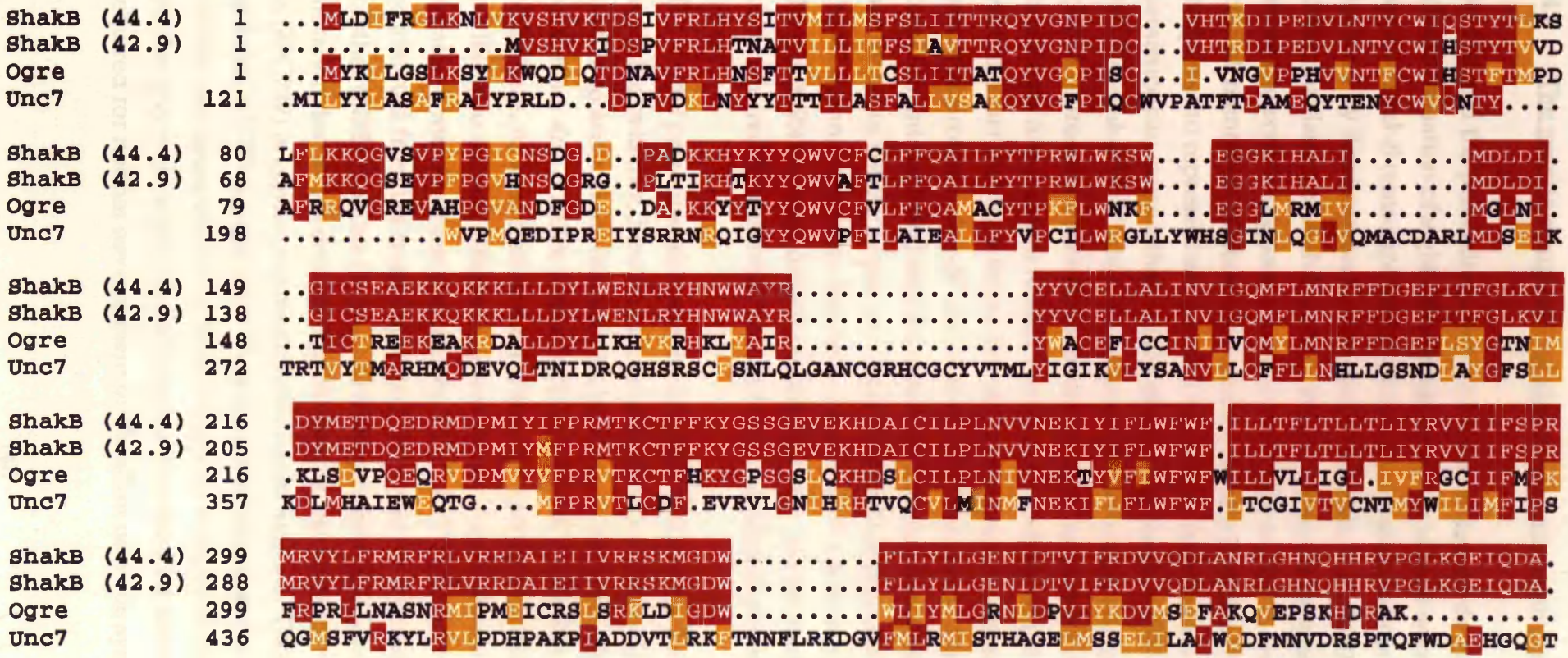


Figure 4.9: Multiple sequence alignment among essential and neural Shak-B proteins, Ogre and Unc-7. Identities (with respect to Shak-B(44.4)) are highlighted in red, conservative changes (A,G; S,T; D,E; K,R; V,I,L,M; N,Q; Y,W,F) in orange.

4.6 The genetic complexity of *shaking-B*: some answers

In this and the preceeding chapter, I have presented details of the characterisation of eight cDNAs representing at least seven alternatively spliced *shak-B* transcript forms. I have also mentioned the P2.4 cDNA form isolated by Krishnan and colleagues (Krishnan, *et al.*, 1993). The positions of different classes of *shak-B* mutant lesion within the proposed coding regions of these transcripts suggests a simple molecular model to account for the complex complementation relationships among *shaking-B* alleles (Figure 4. 10). SIPC8 is proposed to represent an essential transcript, thus lesions within its coding sequence can cause lethal mutations, as seen for $shak\text{-}B^{L41}$ and $shak\text{-}B^{EC201}/shak\text{-}B^{R\text{-}9\text{-}29}$ while, as proposed by Krishnan and colleagues, the P2.4 cDNA represents a *shak-B(neural)* transcript. The $shak\text{-}B^{EC201}$ and $shak\text{-}B^{R\text{-}9\text{-}29}$ mutations complement neither neural nor lethal *shaking-B* alleles, consistent with the fact that their lesion lies within a sequence common to both neural and essential proteins. $shak\text{-}B^{2}$ and $shak\text{-}B^{Passover}$ are neural alleles which complement the lethal *shak-B* mutations, because these *shak-B(neural)* mutations are found in coding sequence unique to the Shak-B(neural) protein encoded by the P2.4 transcript form. Similarly the $shak\text{-}B^{L41}$ allele, which fully complements the giant fibre system defects of *shak-B(neural)* alleles, maps to a translation start required for essential but not for neural Shak-B proteins.

The model presented above is the simplest that can be drawn from the available evidence, however this explanation invokes only two *shak-B* transcript species to account for the complex genetics of the locus while I have reported a further six *shak-B* splice forms (KE2(1.8), P1, SIPC737, SIPC726, SIPC224, and N52) here. At least four of these transcript species (KE2(1.8), P1, SIPC726, N52) are likely to encode different protein forms. Translation of three of these putative proteins (those encoded by KE2(1.8), P1 and SIPC726) is predicted to be initiated from the translation start disrupted by the $shak\text{-}B^{L41}$ lesion, while another (N52) contains the exon G coding region disrupted by the $shak\text{-}B^{EC201}$ mutation, but as yet only the SIPC8 reading frame (present in an identical form in SIPC737) is known to contain *both* of these regions. The fact that $shak\text{-}B^{L41}$ and $shak\text{-}B^{EC201}$ (or $shak\text{-}B^{R\text{-}9\text{-}29}$) fail to complement each other implies that a protein disrupted by *both* of these mutations must be necessary for viability. A similar argument supports the proposal that the P2.4 transcript form is required for normal development of the adult nervous system. However we

cannot yet argue that any identified *shaking-B* splice form or combination thereof is *sufficient* to account for either identified function of *shaking-B*. Whether all *shak-B* transcript forms isolated to date serve specific developmental functions, or whether some may produce nonfunctional proteins and are simply a reflection of alternative splicing as a mechanism of on-off gene control is not yet clear (for further discussion of this issue, see chapter 6). Phenotypic rescue experiments using cDNA constructs alone and in combinations, will be required to resolve this issue[1].

In their genetic analysis of *shak-B*, Baird and colleagues (Baird, *et al.*, 1990) noted that certain chromosomes deficient for loci proximal to the 19E3 region (which harbours the *shak-B(lethal)* genetic function - see Figure 3.2) failed to fully complement the giant fibre system defects of *shak-B* alleles and deficiencies of 19E3. This situation was not due to deletion of a proximal, haploinsufficient locus required for giant fibre system connectivity and neither was it due to deletion of a proximal, haploinsufficient enhancer of *shaking-B*, as very large deletions removing all of the 19E region give wild type flies when heterozygous with balancer chromosomes. In other words, *cis/trans* testing shows a requirement for the 19E3 region and a chromosomal segment proximal to it to be present, intact, in *cis*. The most proximal extent of this proximal region is defined cytologically as the 19E6-7 boundary by the distal breakpoint of the *Df(1)Q539* chromosome, as *Df(1)Q539* fully complements all *shak-B* alleles.

The six deficiencies known to encroach upon the 19E4-6 region show different phenotypic severities when heterozygous with *shak-B(neural)* mutations. The *Df(1)A118*, *Df(1)HC279* and *Df(1)17-351* chromosomes yield the strongest phenotypes. Each of these deficiencies also fails to complement alleles of *R-9-28*, the locus adjacent to *shak-B* on its proximal side. Three other deficiencies, *Df(1)17-481*, *Df(1)A53*, and *Df(1)T2-14A* also fail to fully complement neural mutations of *shak-B*, despite the fact that the distal breakpoints of these deficiencies are, in every case, *proximal* to *R-9-28*. Thus the full *shak-B(neural)* genetic function resides in a large chromosomal region extending from 19E3 perhaps as far proximally as 19E6. This region also contains the *R-9-28* locus (Perrimon, *et al.*, 1989), which is probably unrelated to *shak-B*, yet some of *shak-B(neural)* must reside proximal to *R-9-28*. Thus *R-9-28* is a 'gene within a gene',

---

1 The problems anticipated in achieving phenotypic rescue of the *shak-B* neural mutants do not apply to the rescue of *shak-B* lethality.

an organisation which is unusual in eukaryotes, but which is not without precedent (e.g. Henikoff, *et al.*, 1986).

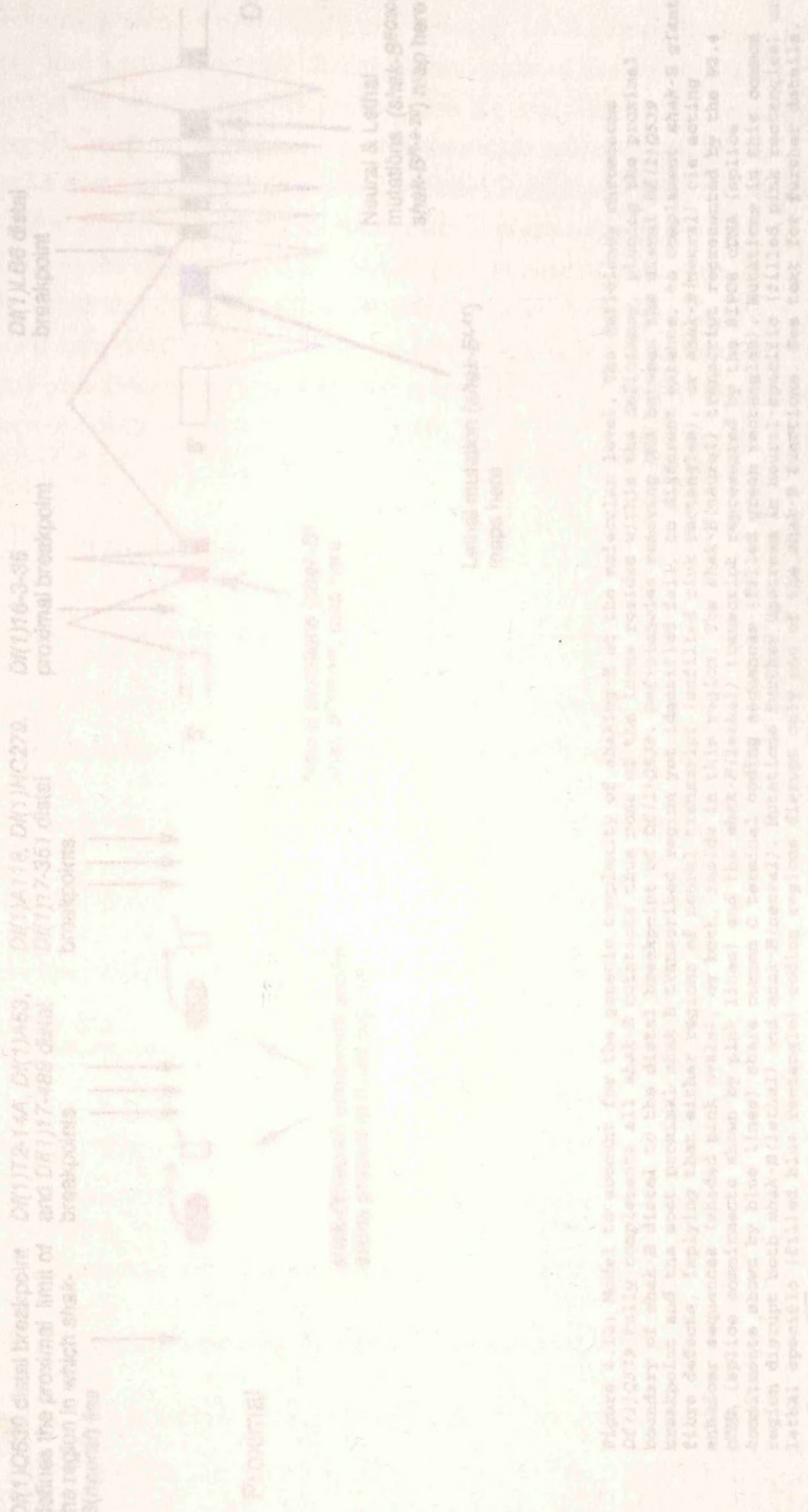
What, then, is the molecular basis of this *cis* dependence on 19E4-6? Broadly speaking three possibilities are apparent (see Figure 4.10). Either (i) some *shak-B cis*-acting enhancer sequences required to drive expression of *shak-B(neural)* transcripts reside in the 19E4-6 region or (ii) some *shak-B(neural)* transcription starts are situated in 19E4-6, or (iii) both of the above are true. The different phenotypic severities of the 6 deficiencies encroaching upon 19E4-6 suggest that either multiple enhancer sequences, or multiple transcription starts, or both are present and that the deficiencies breaking close to 19E3 remove more of these elements than those breaking further proximal.

It should be noted that while some probable transcription starts for *shak-B(lethal)* transcripts have been identified here, no candidate neural transcription start has yet been located, and it is formally possible that neural transcription starts exist upstream in the 19E4-6 region. If this is the case then the *shak-B(neural)* transcription unit is truly vast. The distal breakpoint of *Df(1)T2-14A* has been localised by Alan Griffin (Alan Griffin, unpublished data) within a chromosomal walk called 896. This chromosome walk has yet to unite with the 952 walk containing the identified *shak-B* transcripts but it is nevertheless clear that the *Df(1)T2-14A* breakpoint is at least 110 kb from the most distal exon of P2.4. Whether the *shak-B* transcription unit is, indeed very large, or *shak-B(neural)* function is dependent on an array of distant upstream enhancer sequences, or both, cannot be resolved with the data currently available.

Before any *shaking-B* transcripts had been unequivocally identified, the possibility that *shak-B(neural)* transcripts extended some way proximal to the *Df(1)16-3-35* proximal breakpoint inspired a search for transcripts in the chromosomal walk regions proximal to the *Df(1)16-3-35* breakpoint. These were never found, but a family of cDNA clones describing another, probably unrelated, transcription unit, termed M23, did emerge. M23 has an extensive region of reading frame, but no known homologues. Its sequence will be presented in the appendix to this chapter.

The structural analysis of *shak-B* makes the complex genetics of the locus appear relatively simple in molecular terms (Figure 4.10). I have not yet begun

to address the possible biochemical functions of *shak-B* proteins, and it is to this burning question that the next chapter is devoted.



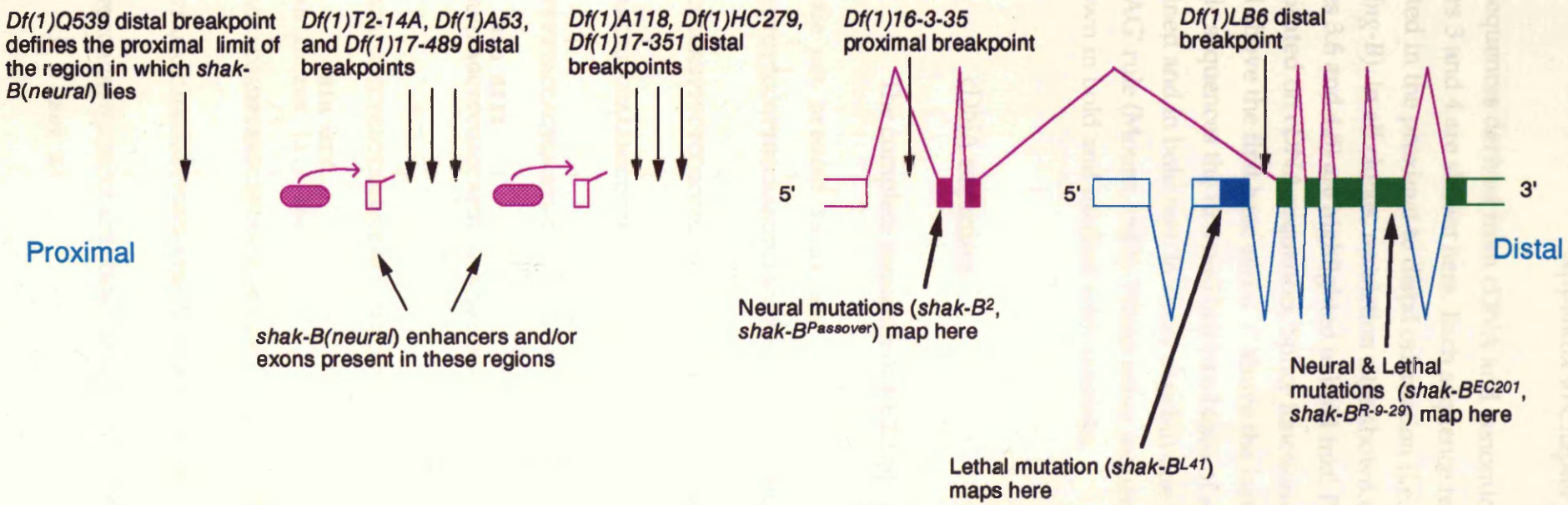


Figure 4.10: Model to account for the genetic complexity of *shaking-B* at the molecular level. The deficiency chromosome *Df(1)Q539* fully complements all *shak-B* mutations thus none of the locus resides within the deficiency, placing the proximal boundary of *shak-B* distal to the distal breakpoint of *Df(1)Q539*. Deficiencies removing DNA between the distal *Df(1)Q539* breakpoint and the most proximal *shak-B* transcribed region yet identified fail, to different extents, to complement *shak-B* giant fibre defects, implying that either regions of neural transcript (unfilled pink rectangles), or *shak-B(neural)* cis acting enhancer sequences (shaded pink ovals), or both, reside in this region. The *shak-B(neural)* transcript represented by the P2.4 cDNA (splice commitments shown by pink lines) and the *shak-B(lethal)* transcript represented by the SIPC8 cDNA (splice commitments shown by blue lines) share common C terminal coding sequences (filled green rectangles). Mutations in this common region disrupt both *shak-B(lethal)* and *shak-B(neural)*. Mutations further upstream in neural-specific (filled pink rectangles) or lethal specific (filled blue rectangle) coding regions disrupt only one of the *shak-B* functions. See text for further details.

DNA sequences derived from cDNA and genomic DNA clones discussed in chapters 3 and 4 are shown here. Each sequence related to *shaking-B* is presented in the proximal to distal orientation (i.e. the direction of transcription of *shaking-B*). In all clones, restriction sites shown on summary diagrams (Figures 3.6 and 4.8) are highlighted in bold text. Proposed translated regions are indicated on cDNA sequences. Splice junctions are indicated with a '\' symbol above the first base and a '/' above the last base of each exon. In genomic sequences the first and last two bases of each intron are shown underlined and in bold text. In every case but one these are found to obey the 'GT to AG' rule (Mount, 1982). Where other sequence features are noted, they are shown in bold and marked with asterisks.

4.A.1 cDNA sequences

4.A.1.a The complete sequence of KE2(1.8)

```
G residue not present in genome                    exon A
*          |          |          |          |          |          60
GCAGTGTCTGTTGCGTTGGCAAAGTGAACGTGCGCGTCTTTTTTCCTGCTCCTGCTGCTC

           |          |          |          |          |          120
CAAAATCCGCAGTTTCTCCGCCTGAACTGATCTCGAAGAAAAGCACGGACAAAAAAAAAA

           |          |          |          |          |          180
ACCTGAAACCGGGGAAAAGCCGTATTCCGATTTCTTTTCCGTAAACGCAAACCCATCGAA

           |          |          |          |          |          240
GTTTTTTTTTTACCAGTGAGCGATCGAAAAATGTGAGGTTGAGGTGCACAGAGTCACCCA

      Hin dIII |          |          |          |          |          300
GAGTCATATAAGCTTCGACCATGTCGGACACGTGACCCGCGCTTATGCCGCGCGATGAAA

                                                   exon A
           |          |          |          |          |          360
CCGATAACACGAGCTGACTAAGCCGATTAGGCCGATAGCAACGATAGCGTCGATAGCCCA

           Hin cII
     AB intron (2.9 kb)                            exon B
           |    /\    |          |          |          |          420
AATCAAACGACGAGTTAACCATTAAAAAAAAGACCAAGCTCAGATAAATCACAAAAAAGT

           |          |          |          |          |          480
GCCAATGTAAAACAGCGGGTTGAGAATACCAGAGAGCGGAGGAACGGACCCGAAGATGCC

           |          |          |          |          |          540
CAGCTCTGCGACAGTCACATCCACTACAACGAACACGCCGCAACCCGAAGAATTGGAAAG

           |  Pst I|  |          |          |          |          600
```

```
CCATCCGCGCCACCTGCAGCACCTGCACCACCCTCATCACCGGAACTACCAGCAACCGCG

         |         |         |         |         |        660
CTTCCGGTTCCGGTTCGAGTCCCGATTAACCGATGCGTCGATAATGGGCCTGCCGCGCAA

         |         |         |         |         |        720
AAGAAAGCACAGCCAGGATTCGGTAGCTCAATCGATCGGTAGCCACACGATACTCTGATA

         |         |         |         |         |        780
TTTTGACACTGATACTATCGATACGCAACGCGAACCAATCGATAACAAACGATAAAAGCA

         |         |         |         |         |        840
AGAAGGGAGGCTAGAGAGAATCATCGCCGATCTTGCCAGCTTTCGCTGACGACAAACCAT
                                                          M

         |         |         |         |         |        900
GTTAGATATATTTCGTGGATTGAAAAACCTTGTAAAGGTCAGTCACGTTAAAACAGATTC
  L   D   I   F   R   G   L   K   N   L   V   K   V   S   H   V   K   T   D   S
Shak-B(14.1) ORF

         |         |         |         |         |        960
GATAGTATTCCGCCTGCATTACTCAATAACTGTGATGATTCTGATGTCATTCTCGCTAAT
  I   V   F   R   L   H   Y   S   I   T   V   M   I   L   M   S   F   S   L   I

         |         |         |         |         |       1020
TATAACCACGCGCCAGTACGTTGGCAATCCGATCGATTGTGTGCACACCAAAGATATTCC
  I   T   T   R   Q   Y   V   G   N   P   I   D   C   V   H   T   K   D   I   P

         |         |      Bam HI       |         |       1080
AGAGGATGTGCTCAACACTTACTGCTGGATCCAGTCCACCTACACGCTCAAGAGCCTCTT
  E   D   V   L   N   T   Y   C   W   I   Q   S   T   Y   T   L   K   S   L   F

         |         |         |         |         |       1140
CCTGAAGAAGCAGGGCGTGTCGGTTCCATATCCGGGCATCGGCAACTCCGACGGTGATCC
  L   K   K   Q   G   V   S   V   P   Y   P   G   I   G   N   S   D   G   D   P

                                                    BC intron (6.7 kb)
         |         |         |         |         |       /\1200
GGCGGACAAGAAGCACTACAAGTACTACCAGTGGGTCTGCTTCTGCCTCTTCTTTCAGCC
  A   D   K   K   H   Y   K   Y   Y   Q   W   V   C   F   C   L   F   F   Q   P

exon C
         |         |         |         |         |       1260
GATATAATGCGGAACAGATGTTGCCTGGACCAATTGCAGCCAAATGCCAACGACAATGGC
  I End

         |    Pst I|         |         |         |       1320
AAATAGCTTGGAACTGCAGGCCCCGAAAAGAAAGAAGTGAAAGGCCAACAAATCCAAACG


         |         |         |         |         |       1380
GAATCGAATGATCGCCGAGATGCACAAGCAACAGCGCGTGTGGGGAAATATTTTAGAGTG

         |         |         |         |         |       1440
GAAAATCGCTGGCCAAGAATTGAGGAACCGCCGCCAGTCGGGCGCCCACTAAACAAACTG

         |         |         |         |         |       1500
AACTTTCACCTCTCAAAACGTAAGACATAGGGATTACCACCTGCCAGGGACTCAGACTCA

         |         |         |         |         |       1560
```

```
AGATCACGATGATGGGGCGACATTGTCATGTGCCAAAGCAAACACACCATATACAATTTG

          |          Pvu II     |          |          |       1620
CACAACTTGCCCAATCCCCAGCTGTCTGTGGATTGTCCGAATCCGCTGCTAATTTACGAT

          |          |          |          |          |       1680
ACAAGTTATTATTAATTTGTCATCGGGTCGGCATTAAAATCCCCCGGCAAAAGAAAATAC

exon C        CD intron (0.79 kb)                    exon D
          |          /\          |          |          |       1740
CGACAACTTATAATGAGCAGGCAATCTTATTTTATACACCAAGATGGCTGTGGAAATCTT

          |          |          |          |          |       1800
GGGAGGGTGGCAAGATTCATGCGCTCATCATGGACTTAGACATAGGCATTTGTTCCGAAG

  Poly(A) sequence probably derived from internal priming
  **************   |          |          |          |       1860
CCGAAAAAAAAAAAAAA
```

4.A.1.b Sequence of P1 cDNA (incomplete)

No introns are known to be present within this sequence

```
          | Acc I    |          |          |          |         60
GCCACGTGTGCGTATACGCAATTTTCACGGTGTGTGCGTGGACTTTTCTTTTTTGAGTAC

          |          |          |          |          |        120
ACAGAAGAAAATATAGTTCTAGTCTAGTGCTAAATGCAGAAACAATATTTTACAAGAACG

          |          |          |          |          |        180
TTTTCTTCTAATTAGATATTAAAGTATTAATCTAATATGGTTGCTATGTGTCACTAATAT

          |          |          |          |          |        240
TTTACACATGATTTATGAATTATAACCACTTCATTGTCAACTATTCTTGTGTCTCAATAT

          |          |          |          |          |        300
TTCGCTTGCTTTGCTATCGCTGTGTTGATTATATCGCGCCGCGGAAGTGTTAACATGCAT

          |          |          |          |          |        360
TAATTCAGTCATGACGTG......Up to 50 bases not determined.......

          |          |          |          |          |        420
TTGTTTTGATTTATTATCGGCAGCTCGGCGGGAAAGTTTGTTTATAAACATTGCAATTTT

          |          |          |          |          |        480
AAAATTTACAGTTCATTTGGAAACAGTTGTTCTCCAGTTTTTGGTCTTCCATTTTTCCCC

          |          |          |          |          |        540
TACGTCAGTCAGCATAAGTTTATTAACATATGTGAAGGAGGGGGTGATGCCACATTTTAT

          |          |          |          |          |        600
TTATATTCTCGAATTAGTTTTTTTATTTGTTTTTCTTTGCGCAATTTGCAATGGTATTTA

          |          |          |          |          |        660
AGCATAAAAAAATTATTTTAATACCGAGAAAAGTGCTAGCACTTGATTTCCTATATAGTC
```

```
                                                                  720
TAGTTTCCTTCGTGTTGTTTTAGACTAAAATACACAATATAATAATATTAAACGTGAGCC

                                                                  780
AAAAATAAATATGTTAGTAAAGAATGAGGAGGTAAAAATGTTCGTTTTTCATTCAGAAAA

                                                                  840
AAAGTTATCTTTTCACAATGATTGATCAATGGGCATGTGTACAACAATGAGGAAAGCACA

                                                                  900
TGTGATTTCATAAAATAGTGTTTCCGCTAATAAACTTAAAAATTCGTTATGCTAAATTTC

          Bgl II                                                  960
AAGCAGGCGAGAGATCTTGAAAAGTCCCATAATGCATGCAGTGAGCTTTGTGAAAATATA

                                                                 1020
GAAGTGAAAGCGCGAAAAATTGTGAAATCAAAATTGTTCATCTCATTGTCAGCAGTAACC

                                                                 1080
ATTAAAAAAAAAGACCAAGCTCAGATAAATCACAAAAAAGTGCCAATGTAAAACAGCGGGT

                                                                 1140
TGAGAATACCAGAGAGCGGAGGAACGGACCCGAAGATGCCCAGCTCTGCGACAGTCACAT
                                                        Pst I
                                                                 1200
CCACTACAACGAACACGCCGCAACCCGAAGAATTGGAAAGCCATCCGCGCCACCTGCAGC

                                                                 1260
ACCTGCACCACCCTCATCACCGGAACTACCAGCAACCGCGCTTCCGGTTCCGGTTCGAGT

                                                                 1320
CCCGATTAACCGATGCGTCGATAATGGGCCTGCCGCGCAAAAGAAAGCACAGCCAGGATT

                                                                 1380
CGGTAGCTCAATCGATCGGTAGCCACACGATACTCTGATATTTTGACACTGATACTATCG

                                                                 1440
ATACGCAACGCGAACCAATCGATAACAAACGATAAAGCAAGAAGGGAGGCTAGAGAGATC

                                                                 1500
ATCGCCGATCTTGCCAGCTTTCGCTGACGACAAACCATGTTAGATATATTTCGTGGATTG
                                        M  L  D  I  F  R  G  L
                                        Shak-B(14.0) ORF

                                                                 1560
AAAAACCTTGTAAAGGTCAGTCACGTTAAAACAGATTCGATAGTATTCCGCCTGCATTAC
 K  N  L  V  K  V  S  H  V  K  T  D  S  I  V  F  R  L  H  Y

                                                                 1620
TCAATAACTGTGATGATTCTGATGTCATTCTCGCTAATTATAACCACGCGCCAGTACGTT
 S  I  T  V  M  I  L  M  S  F  S  L  I  I  T  T  R  Q  Y  V

                                                                 1680
GGCAATCCGATCGATTGTGTGCACACCAAAGATATTCCAGAGGATGTGCTCAACACTTAC
 G  N  P  I  D  C  V  H  T  K  D  I  P  E  D  V  L  N  T  Y

   Bam HI                                                        1740
TGCTGGATCCAGTCCACCTACACGCTCAAGAGCCTCTTCCTGAAGAAGCAGGGCGTGTCG
 C  W  I  Q  S  T  Y  T  L  K  S  L  F  L  K  K  Q  G  V  S

                                                                 1800
GTTCCATATCCGGGCATCGGCAACTCCGACGGTGATCCGGCGGACAAGAAGCACTACAAG
```

```
 V  P  Y  P  G  I  G  N  S  D  G  D  P  A  D  K  K  H  Y  K

         |        |        |        |        |      1860
TACTACCAGTGGGTCTGCTTCTGCCTCTTCTTTCAGGTATAAGAAGGTTGTGTCTCCGGA
 Y  Y  Q  W  V  C  F  C  L  F  F  Q  V  End

         |        |        |        |        |      1920
CCACTTGAAACATTAATTAATTTCAAAATATCGACACCGATAGTCGTTGTCGAATCATTG

Kpn I    |        |        |        |        |      1980
GTACCAAGCTAACAAATTTTCATCCATGATAAATTTGTATTTAATTTACTTTCCAATAAC

         |        |        |        |        |      2040
TATCACTTTGATATAAACGTGTAAATGGGAATAGCTCAAACTGACAGTATTAAGTCACTT

         |        |        |        |        |      2100
CTGCATCATTTAAAGCATCACAGTTGAAAGCGTTTTCAAGAAATTAAAGTGTCAAGCGAA

         |        |        |        |        |      2160
TAAGGGAATCTATTGGCCAACGGTAATTAAGCATCTCTGTTTTTTAACGCAATATTAATT

         |        |        |        |        |      2220
AAAATTCGACGCCATAGCTCATCCTCTTAAAAGGCTACGCTGCGCACGGATTAAGCCAGA

  Acc I  |        |        |        |        |      2280
AAGTCTACTGAGAAATCAGCCATTTATCAGTGGTTCGTTATGTTTGGGGGCGTTGAATGA

         |        |        |        |        |      2340
TTTGATGGATTAAGTAATTAAGCGATTTCTCGCCGACTGAGCGCACCATTGGGATCTATA

         |        |        |        |        |      2400
CATACTTTCAATATCCAAATATTTGTTCCACTTATTTCCGATTTCTTTTTTTTTATTTTG

         |        |        |        |        |      2460
TTTTGCGGTTTTGGCCTTTTGTGAGCCAAAATGAAATTGTTTAAGTATTTCGTGCGTGTT

         |        |        |        |        |      2520
TTGTTATGTTGCTGCTGAATGGGCTGGGCCCTTGTATCGCTGCGAGTATCCAAAGATAGT

         |        |        |        |        |      2580
CGATCTATGACGTAGGGGATGTCAGGCAGAGACGGGGATACTGGGCGAAAAAGAGAGGGG

         |        |        |        |        |      2640
GATAGAGAGAGAAAGCGACAATCTAAAGATATTATAATAAGAAGATGNGTAACGGCCGGG

         |        |        |        |        |      2700
ATGCACTACTGGCTGAATGATACCTTATTGTGGCCAAAATAAATAACTTCAAAGCGGAGA

         |        |        |        |        |      2760
```

```
CAGATATTTTTTTTTTTGTGTTTTTTGTATTGGGTTACACAAGCTCATATCATAATTTTG

                                                        Acc I
         |         |         |         |         |      2820
GGAGTGGCACCATCCCTTACACATTATTATACCCGTTACTCGTAGATTGAAAGGGTATAC

         |         |         |         |         |      2880
TAGATTTGTTGAAAAGTATGTAACAGGCAGAAGTAAGCGTTTCCGACTATATAAAGTATA

         |         |         |         |         |      2940
TATATTCTTGATCAGGATCAATATCCGAGTCAATCCAGCCGTCTGTCCGTATGAACACCG

 Bgl II  |         |         |         |         |      3000
TAGATCTCAGGATCTATAAAAGCATGAAGGGTGAGCTCAGCATTAGAAGTAGAGACAAAG

         |         |         |         |         |      3060
ACATAGGCACAAGTTTGTTGACCAATGTTGCCATGCCCACAAACCGCAACAAAACTTCCA

         |         |         |         |         |      3120
CGTCCATGTTTTGAACCATTTTTCGATATTTTTTTAAAATTTTATTAGTCTTCTAAACTT

         |         |         |         |         |      3180
ATCTGATTTGCCAAAACATTTTTGCCACGCCCACTCGTATGCCCTAAAGCCGACAAACCG

         |         |         |         |         |      3240
GTCACGCCCACACTTTAAAACAACATTTTAATTTTTTGTCATATTATTCCCCAGAATCTA

  Eco RV |         |         |         |         |      3300
TCGATATCCCAGAAAAATTATAAAATTTCGCACTCGCATTCACACTAGCTGAGTAACGGG

                                                        Acc I
         |         |         |         |         |      3360
TATCTGATAGTCGGGAAACTCGACTATAGCATTATTTTTTTTGCATTAACTTACAGTAGA

         |         |         |         |         |      3420
CATTATCTATGAGAATCTCAAAAATTTTGGTCGTGGCACCGCCCCTTAAATTTAATTAAT

         |         |         |         |         |      3480
TCATTTTCGTATAGGGATTTCTCTAATTTCGATGCTCTTCTACATACATATCATATCATA

         |         |         |         |         |      3540
ATCATACATATGATTACAGTATTAAATAATTTTGACAAATAATAACCACAAAAATATTGA

         |         |         |         |         |      3600
AACAATTCAGCAACAGCCTTGTTTAACTTGTTTAGCTAGTACTCAAATTTGAAAGTAGGT

         |         |         |         |         |      3660
AGGTCCTAAAAACTAAAATTTCTCCATAAATATTGCATTCAAATTAAGAAAAATTTAATT
```

```
          |  Acc I  |         |         |         |       3720
ACCTACAACACAGTATACATATTATGGTGAAACTATGTGAAACTCAGTTTATAAACAAAT

          |         |         |         |         |       3780
AAATAAAAATAAATAAATAAATAAATATAAGTTTTTGTTTAAACACCTAAAAGCGCTATT

                                    Poly (A) tail
          |         |         |         |         |       3840
TACAACACTCAATTACAGGGACAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
```

4.A.1.c The complete sequence of SIPC8 (incorporating the B10 cDNA fragment)

Bases in lower case are not present in SIPC8, or in any other SIPC cDNA clones, as they lie between the divergent primers used in the cloning of these cDNAs. The lower case sequence is inferred from the sequence in this region present in KE2(1.8), P1, and psBGC.

```
G residue not present in genome                    exon A
*         |         |         |         |         |       60
GAGTGTCTGTTGCGTTGGCAAAGTGAACGTGCGCGTCTTTTTTCCTGCTCCTGCTGCTCC

          |         |         |         |         |      120
AAAATCCGCAGTTTCTCCGCCTGAACTGATCTCGAAGAAAAGCACGGACAAAAAAAAAAA

          |         |         |         |         |      180
CCTGAAACCGGGGAAAAGCCGTATTCCGATTTCTTTTCCGTAAACGCAAACCCATCGAAG

          |         |         |         |         |      240
TTTTTTTTTTACCAGTGAGCGATCGAAAAATGTGAGGTTGAGGTGCACAGAGTCACCCAG

        Hin dIII  |         |         |         |         |      300
AGTCATATAAGCTTCGACCATGTCGGACACGTGACCCGCGCTTATGCCGCGCGATGAAAC

          |         |         |         |         |      360
CGATAACACGAGCTGACTAAGCCGATTAGGCCGATAGCAACGATAGCGTCGATAGCCCAA

         AB intron (2.9 kb)
exon A       Hin cII              exon B
          |   /\    |         |         |         |      420
ATCAAACGACGAGTTAACCATTAAAAAAAAGACCAAGCTCAGATAAATCACAAAAAAGTG

          |         |         |         |         |      480
CCAATGTAAAACAGCGGGTTGAGAATACCAGAGAGCGGAGGAACGGACCCGAAGatgccc

          |         |         |         |         |      540
agctctgcgacagtcacatccactacaacgaacacgccgcaacccgaagaattggaaagc

          |  Pst I  |         |         |         |      600
catccgcgccacctgcagcacctgcaccaccctcatcaccggaactaccagcaaccgcgc

          |         |         |         |         |      660
ttccggttccggtTCGAGTCCCGATTAACCGATGCGTCGATAATGGGCCTGCCGCGCAAA

          |         |         |         |         |      720
AGAAAGCACAGCCAGGATTCGGTAGCTCAATCGATCGGTAGCCACACGATACTCTGATAT

          |         |         |         |         |      780
```

```
TTTGACACTGATACTATCGATACGCAACGCGAACCAATCGATAACAAACGATAAAAGCAA

          |         |         |         |         |       840
GAAGGGAGGCTAGAGAGAATCATCGCCGATCTTGCCAGCTTTCGCTGACGACAAACCATG
                                                      M

          |         |         |         |         |       900
TTAGATATATTTCGTGGATTGAAAAACCTTGTAAAGGTCAGTCACGTTAAAACAGATTCG
 L  D  I  F  R  G  L  K  N  L  V  K  V  S  H  V  K  T  D  S
Shak-B(44.4) ORF

          |         |         |         |         |       960
ATAGTATTCCGCCTGCATTACTCAATAACTGTGATGATTCTGATGTCATTCTCGCTAATT
 I  V  F  R  L  H  Y  S  I  T  V  M  I  L  M  S  F  S  L  I

          |         |         |         |         |      1020
ATAACCACGCGCCAGTACGTTGGCAATCCGATCGATTGTGTGCACACCAAAGATATTCCA
 I  T  T  R  Q  Y  V  G  N  P  I  D  C  V  H  T  K  D  I  P

          |         |  Bam HI |         |         |      1080
GAGGATGTGCTCAACACTTACTGCTGGATCCAGTCCACCTACACGCTCAAGAGCCTCTTC
 E  D  V  L  N  T  Y  C  W  I  Q  S  T  Y  T  L  K  S  L  F

          |         |         |         |         |      1140
CTGAAGAAGCAGGGCGTGTCGGTTCCATATCCGGGCATCGGCAACTCCGACGGTGATCCG
 L  K  K  Q  G  V  S  V  P  Y  P  G  I  G  N  S  D  G  D  P

          |         |         |         |         |      1200

                                                  BD intron (8.0 kb)
                                                      exon B  /\
GCGGACAAGAAGCACTACAAGTACTACCAGTGGGTCTGCTTCTGCCTCTTCTTTCAGGCA
 A  D  K  K  H  Y  K  Y  Y  Q  W  V  C  F  C  L  F  F  Q  A

exon D    |         |         |         |         |      1260
ATCTTATTTTATACACCAAGATGGCTGTGGAAATCTTGGGAGGGTGGCAAGATTCATGCG
 I  L  F  Y  T  P  R  W  L  W  K  S  W  E  G  G  K  I  H  A

          |         |         |         |         |      1320
CTCATCATGGACTTAGACATAGGCATTTGTTCCGAAGCCGAGAAAAAACAAAAAAAGAAA
 L  I  M  D  L  D  I  G  I  C  S  E  A  E  K  K  Q  K  K  K

                              DE intron (0.53 kb)
                                  Eco RV
          |         | exon D  /\    exon E      |      1380
TTACTTTTAGATTATTTGTGGGAAAATTTAAGATATCACAATTGGTGGGCGTACAGATAT
 L  L  L  D  Y  L  W  E  N  L  R  Y  H  N  W  W  A  Y  R  Y

                                 EF intron (2.0 kb)
          |         |       exon E  /\  exon F  |      1440
TACGTGTGTGAGCTGCTCGCCCTTATAAATGTGATAGGTCAAATGTTTCTTATGAATCGA
 Y  V  C  E  L  L  A  L  I  N  V  I  G  Q  M  F  L  M  N  R

          |         |         |         |         |      1500
TTTTTCGATGGCGAATTTATAACATTTGGCCTGAAAGTGATAGATTATATGGAGACCGAT
 F  F  D  G  E  F  I  T  F  G  L  K  V  I  D  Y  M  E  T  D

          | Bam HI  |         |         |         |      1560
CAGGAAGATCGCATGGATCCGATGATTTACATATTTCCCAGAATGACCAAATGTACATTT
 Q  E  D  R  M  D  P  M  I  Y  I  F  P  R  M  T  K  C  T  F

                              FG intron (61 bases)
          |       exon F /\  exon G     |         |      1620
```

```
TTTAAATATGGTTCCAGTGGGGAGGTGGAGAAACACGACGCCATTTGCATTTTACCATTA
 F  K  Y  G  S  S  G  E  V  E  K  H  D  A  I  C  I  L  P  L

         |        |        |        |        |      1680
AACGTTGTTAATGAGAAGATTTACATTTTCCTTTGGTTTTGGTTTATATTATTAACGTTT
 N  V  V  N  E  K  I  Y  I  F  L  W  F  W  F  I  L  L  T  F
```

The Hin cII below is not present in the genomic sequence: the starred G residue is polymorphic, an A being present in the genomic sequence (psBGK- see below). This polymorphism is a silent, third base substitution.

```
        Hin cII
        *|       |        |        |        |      1740
CTCACATTGTTAACGCTAATATACAGGGTGGTTATTATATTCTCTCCTCGAATGAGGGTC
 L  T  L  L  T  L  I  Y  R  V  V  I  I  F  S  P  R  M  R  V

                    GH intron (0.39 kb)
Acc I    |    exon G  /\  exon H       |        |      1800
TACTTATTTCGTATGCGATTTAGGTTAGTGCGTCGTGACGCTATTGAAATAATCGTTCGT
 Y  L  F  R  M  R  F  R  L  V  R  R  D  A  I  E  I  I  V  R

         |        |        |        |        |      1860
CGTTCAAAGATGGGCGATTGGTTTTTGTTGTATTTACTAGGTGAAAACATAGATACAGTT
 R  S  K  M  G  D  W  F  L  L  Y  L  L  G  E  N  I  D  T  V

         |        |        |        |        |      1920
ATATTTCGTGATGTTGTACAGGACTTAGCGAATCGTTTAGGACATAACCAACACCACAGG
 I  F  R  D  V  V  Q  D  L  A  N  R  L  G  H  N  Q  H  H  R

         |        |        |        |        |      1980
GTGCCTGGCTTAAAAGGTGAAATACAGGATGCATGATATTGGGAGTATTAGAAACAATAC
 V  P  G  L  K  G  E  I  Q  D  A End

         |        |        |     Eco RI      |      2040
AAAATGCAATTTGTCGTCTCCATTTAAAAACCATCGAATTCGATAACAAAAATGTCGAAA

         |        |        |        | Xba I  |      2100
GCAAGAAAAAGATTAAGAAGGACAATTACAACCACAAAGGAATCTAGAGATCTTCGCAGC
                                      Bgl II

         |        |        |        |        |      2160
AGCCGCTTCATTAAAACTTACAACTCAACACACCGCTAAAAAATCTTTAAAAAAAAAAA
```

4.A.1.d SIPC726 sequence (incomplete)

Bases in lower case have not been sequenced from this clone; instead the internal sequence (lower case) is inferred, from restriction mapping data (not shown) to be identical to that in SIPC8 and SIPC737.

```
exon B'
         |        |        |        | Bgl II         60
TAATAAACTTAAAAATTCGTTATGCTAAATTTCAAGCAGGCGAGAGATCTTGAAAAGTCC

         |        |        |        |        |      120
CATAATGCATGCAGTGAGCTTTGTGAAAATATAGAAGTGAAAGCGCGAAAAATTGTGAAA

         |        |        |        |        |      180
TCAAAATTGTTCATCTCATTGCCAGCAGTAACCATTAAAAAAAAGACCAAGCTCAGATAA
```

```
        |         |         |         |         |       240
ATCACAAAAAAGTGCCaatgtaaaacagcgggttgagaataccagagagcggaggaacgg

        |         |         |         |         |       300
acccgaagatgcccagctctgcgacagtcacatccactacaacgaacacgccgcaacccg

        |         |      Pst I        |         |       360
aagaattggaaagccatccgcgccac ctgcag cacctgcaccaccctcatcaccggaact

        |         |         |         |         |       420
accagcaaccgcgcttccggttccggttcgagtcccgattaaccgatgcgtcgataatgg

        |         |         |         |         |       480
gcctgccgcgcaaaagaaagcacagccaggattcggtagctcaatcgatcggtagccaca

        |         |         |         |         |       540
cgatactctgatattttgacactgatactatcgatacgcaacgcgaaccaatcgataaca

        |         |         |         |         |       600
aacgataaaagcaagaagggaggctagagagaatcatcgccgatcttgccagctttcgct

        |         |         |         |         |       660
gacgacaaaccatgttagatatatttcgtggattgaaaaaccttgtaaaggtcagtcacg
           m  l  d  i  f  r  g  l  k  n  l  v  k  v  s  h  v

        |         |         |         |         |       720
ttaaaacagattcgatagtattccgcctgcattactcaataactgtgatgattctgatgt
   k  t  d  s  i  v  f  r  l  h  y  s  i  t  v  m  i  l  m  s

        |         |         |         |         |       780
cattctcgctaattataaccacgcgccagtacgttggcaatccgatcgattgtgtgcaca
   f  s  l  i  i  t  t  r  q  y  v  g  n  p  i  d  c  v  h  t

        |         |         |      Bam HI       |       840
ccaaagatattccagaggatgtgctcaacacttactgct ggatcc agtccacctacacgc
   k  d  i  p  e  d  v  l  n  t  y  c  w  i  q  s  t  y  t  l

        |         |         |         |         |       900
tcaagagcctcttcctgaagaagcagggcgtgtcggttccatatccgggcatcggcaact
   k  s  l  f  l  k  k  q  g  v  s  v  p  y  p  g  i  g  n  s

        |         |         |         |         |       960
ccgacggtgatccggcggacaagaagcactacaagtactaccagtgggtctgcttctgcc
   d  g  d  p  a  d  k  k  h  y  k  y  y  q  w  v  c  f  c  l

   B'D intron (8 kb)
  exon B' /\  exon D    |         |         |      1020
tcttctttcaggcaatcttattttatacaccaagatggctgtggaaaTCTTGGGAGGGTG
   f  f  q  a  i  l  f  y  t  p  r  w  l  w  k  S  W  E  G  G

        |         |         |         |         |      1080
GCAAGATTCATGCGCTCATCATGGACTTAGACATAGGCATTTGTTCCGAAGCCGAGAAAA
   K  I  H  A  L  I  M  D  L  D  I  G  I  C  S  E  A  E  K  K

                                  DE' intron (0.53 kb)
        |         |         |         |      Eco RV      1140
                                 exon D  /\  exon E'
AACAAAAAAAAGAAATTACTTTTAGATTATTTGTGGGAAAATTTAA GATATC ACAATTGGT
   Q  K  K  K  L  L  L  D  Y  L  W  E  N  L  R  Y  H  N  W  W

        |         |         |         |         |      1200
GGGCGTACAGATATTACGTGTGTGAGCTGCTCGCCCTTATAAATGTGATAGGATAGGTGN
   A  Y  R  Y  Y  V  C  E  L  L  A  L  I  N  V  I  G End
```

```
      Poly(A) sequence probably derived from internal priming
         |          ***************      |          |      1260
GTACTAGTGCACGGCAAACAAAAAAAAAAAAAA
```

4.A.2 Genomic DNA sequences

4.A.2.a Sequence of psBGG (incomplete)

The sequence of approximately 1kb at the proximal end of this subclone has not been determined.

```
         |          |          |          |          |      1059
ACAGCGCTCCTCGCACACACCATTTCAGACTCTGTTCCCCCACCGCGACTGGCCACCGCT

         |          |          |          |          |      1119
TTTCCCACCGAGTATGCGTGCCTCGCGAGCAGGGAACGCGTTGCTGCCTCCCAAAAATTG

         |  Bgl II  |          |          |          |      1179
TTGGCAGCACAGAGATCTGCTCTCTGGCTTTTCCCCGCTCTCTGTCACTGTCTGGATTTC

         |          |          |          |          |      1239
CCTCTCATTCTCCCTCTGTCTCGCTGTCTCTCTGCCTGTCAGCTGTGGGCAGGCCATTTT

            First base of KE2(1.8) (underlined)
            First base of SIPC8 (bold)
         |          |          |          |          |      1299
CCGAGTGGCTTTTCTTGGCCAGTGCCCGCCAGTGTCTGTTGCGTTGGCAAAGTGAACGTG

exon A
         |          |          |          |          |      1359
CGCGTCTTTTTTCCTGCTCCTGCTGCTCCAAAATCCGCAGTTTCTCCGCCTGAACTGATC

         |          |          |          |          |      1419
TCGAAGAAAAGCACGGACAAAAAAAAAAACCTGAAACCGGGGAAAAGCCGTATTCCGATT

         |          |          |          |          |      1479
TCTTTTCCGTAAACGCAAACCCATCGAAGTTTTTTTTTTACCAGTGAGCGATCGAAAAAT

         |          |          |          |  Hin dIII 1539
GTGAGGTTGAGGTGCACAGAGTCACCCAGAGTCACCCAGAGTCATATAAGCTT
```

4.A.2.b Complete sequence of psBGB

```
                                                A in bold = first
                                                base of SIPC224
Hin dIII |  exon A  |          |          |          |        60
AAGCTTCGACCATGTCGGACACGTGACCCGCGCTTATGCCGCGCGATGAAACCGATAACA

         |          |          |          |          |      120
CGAGCTGACTAAGCCGATTAGGCCGATAGCAACGATAGCGTCGATAGCCCAAATCAAACG

            AB intron (included within exon A')
         |          |          |          |          |      180
ACGAGTGTGAGAATTAAATTCGCGCAACTAGTGTGCAAACAAACACCGAGCGTTTTCTAT
```

```
        |         |         |         |         |       240
ATTTTCACAACACCCACGAAAGATCATTGACTGCAATGCGGGATCATAACATATAGTAGG

        |         |         |         |         |       300
AAGCCACATGCCCACAGAATCAGAAATTATATTTATATATACACGTATATATAGTATTGT

        |         |         |         |         |       360
AGGGGCTGTGTGCAGGCCACCGCTTGATCATAATTAACTCTTACAAAGTAAATAAACCAA

   Eco RI |       |         |         |         |       420
CTGCGAATTC
```

4.A.2.c Sequence of psBGP (incomplete)

```
Eco RI  |         |         |         |         |        60
GAATTCGGGGGCCAGTGTGTTGCCCCCGAAACCGAAAACCAAAAACCGAAAACAAACAGG

        |         |         |         |         |       120
CGATCCACAAATTTTTACAATTAACATTCGACACGATTGTCCAAAAACCAAAATGGAAAC

                                        exon A'     intron
        |         |         |         |        |/       180
CGAAAAAACCAAAATACAAATTTTATTTTTCATTTTATTTTTATATTTTATGTGAGTTTG

        |         |         |         |         |       240
TTGTCGTCTTATCATTCGCACATAATTGGGCATGAAACATAAAGTTTGAGTAACTTTTAG

        |         |         |         |         |       300
AGTAACTTTTGCTTTAATTTTGTTA.....Remainder of clone (approximately
                              1.2 kb) not sequenced
```

4.A.2.d Complete sequence of psBGC

```
Bgl II  |         |         |         |         |        60
AGATCTTGAAAAGTCCCATAATGCATGCAGTGAGCTTTGTGAAAATATAGAAGTGAAAGC

                        AG (underlined) = last bases of
                        first intron of SIPC737;
                        AG (bold) = last bases of first
                        intron in SIPC8, KE2(1.8), SIPC224.
        |         |         |         | \  \    |       120
GCGAAAAATTGTGAAATCAAAATTGTTCATCTCATTGTCAGCAGTAACCATTAAAAAAAA

        |         |         |         |         |       180
GACCAAGCTCAGATAAATCACAAAAAAGTGCCAATGTAAAACAGCGGGTTGAGAATACCA

        |         |         |         |         |       240
GAGAGCGGAGGAACGGACCCGAAGATGCCCAGCTCTGCGACAGTCACATCCACTACAACG

        |         |         |         | Pst I   |       300
AACACGCCGCAACCCGAAGAATTGGAAAGCCATCCGCGCCACCTGCAGCACCTGCACCAC

        |         |         |         |         |       360
CCTCATCACCGGAACTACCAGCAACCGCGCTTCCGGTTCCGGTTCGAGTCCCGATTAACC

        |         |         |         |         |       420
```

```
GATGCGTCGATAATGGGCCTGCCGCGCAAAAGAAAGCACAGCCAGGATTCGGTAGCTCAA

                                                          480
TCGATCGGTAGCCACACGATACTCTGATATTTTGACACTGATACTATCGATACGCAACGC

                                                          540
GAACCAATCGATAACAAACGATAAAGCAAGAAGGGAGGCTAGAGAGATCATCGCCGATCT

                                                          600
TGCCAGCTTTCGCTGACGACAAACCATGTTAGATATATTTCGTGGATTGAAAAACCTTGT

                                                          660
AAAGGTCAGTCACGTTAAAACAGATTCGATAGTATTCCGCCTGCATTACTCAATAACTGT

                                                          720
GATGATTCTGATGTCATTCTCGCTAATTATAACCACGCGCCAGTACGTTGGCAATCCGAT
                                                    Bam HI
                                                          780
CGATTGTGTGCACACCAAAGATATTCCAGAGGATGTGCTCAACACTTACTGCTGGATCC
```

4.A.2.e Complete sequence of psBGD

```
Bam HI                                                    60
GGATCCAGTCCACCTACACGCTCAAGAGCCTCTTCCTGAAGAAGCAGGGCGTGTCGGTTC

                                                          120
CATATCCGGGCATCGGCAACTCCGACGGTGATCCGGCGGACAAGAAGCACTACAAGTACT

                          exon B                    intron
                                                  /       180
ACCAGTGGGTCTGCTTCTGCCTCTTCTTTCAGGTATAAGAAGGTTGTGTCTCCGGACCAC

                                                    Kpn I
                                                          240
TTGAAACATTAATTAATTTCAAAATATCGACACCGATAGTCGTTGTCGAATCATTGGTAC

                                                          300
CAAGCTAACAAATTTTCATCCATGATAAATTTGTATTTAATTTACTTTCCAATAACTATC

                                                          360
ACTTTGATATAAACGTGTAAATGGGAATAGCTCAAACTGACAGTATTAAGTCACTTCTGC

                                                          420
ATCATTTAAAGCATCACAGTTGAAAGCGTTTTCAAGAAATTAAAGTGTCAAGCGAATAAG

                                                          480
GGAATCTATTGGCCAACGGTAATTAAGCATCTCTGTTTTTTAACGCAATATTAATTAAAA

                                                          540
TTCGACGCCATAGCTCATCCTCTTAAAAGGCTACGCTGCGCACGGATTAAGCCAGAAAGT

Acc I                                                     600
CTACTGAGAAATCAGCCATTTATCAGTGGTTCGTTATGTTTGGGGGCGTTGAATGATTTG

                                                          660
ATGGATTAAGTAATTAAGCGATTTCTCGCCGACTGAGCGCACCATTGGGATCTATACATA

                                                          720
CTTTCAATATCCAAATATTTGTTCCACTTATTTCCGATTTCTTTTTTTTTATTTTGTTTT
```

```
          |         |         |         |         |        780
GCGGTTTTGGCCTTTTGTGAGCCAAAATGAAATTGTTTAAGTATTTCGTGCGTGTTTTGT

          |         |         |         |         |        840
TATGTTGCTGCTGAATGGGCTGGGCCCTTGTATCGCTGCGAGTATCCAAAGATAGTCGAT

          |         |         |         |         |        900
CTATGACGTAGGGGATGTCAGGCAGAGACGGGGATACTGGGCGAAAAAGAGAGGGGGATA

          |         |         |         |         |        960
GAGAGAGAAAGCGACAATCTAAAGATATTATAATAAGAAGATGNGTAACGGCCGGGATGC

          |         |         |         |         |       1020
ACTACTGGCTGAATGATACCTTATTGTGGCCAAAATAAATAACTTCAAAGCGGAGACAGA

          |         |         |         |         |       1080
TATTTTTTTTTTTGTGTTTTTTGTATTGGGTTACACAAGCTCATATCATAATTTTGGGAG

          |         |         |         |        |Acc I  1140
TGGCACCATCCCTTACACATTATTATACCCGTTACTCGTAGATTGAAAGGGTATACTAGA

          |         |         |         |         |       1200
TTTGTTGAAAAGTATGTAACAGGCAGAAGTAAGCGTTTCCGACTATATAAAGTATATATA

          |         |         |         |         |       1260
TTCTTGATCAGGATCAATATCCGAGTCAATCCAGCCGTCTGTCCGTATGAACACCGTAGA

Bgl II    |         |         |         |         |       1320
TCT
```

4.A.2.f Sequence of psBGE (incomplete)

A few bases adjacent to the EcoRI sites at each end of this clone have not been determined.

```
          |         |         |         |         |         60
CTTCGGAGACGATTCTTTTCTCTTTTAGCCTTTCGCTTTGATTGTTGCTTCTGTTGCAGT

          |         |         |         |         |        120
TGCATTTACGCGATGCTTTTGCGATTTGAACGCTTGCCAAATTGCTTTCCACACACGCAC

          |         |         |         |         |        180
AGAGGAAAGTGGAAAAGAGGGTTGGAAAAGGGTGGTGGGGAGGCCACCCTTCTTTTCTCT

          |         |         |         | Acc I   |        240
TCCTTATTCGTGTTATCGTCGGGGAAATGCAATTGGCGATTGTATACACAAAAGGGCAGA

          |         |         |         |         |        300
CTGAGATGATGGAGGCCAAAGCAAATCAAACGCATCATTTGACTTCCGGCAGGGCGTGGC

          |         |         |         |         |        360
AACATACCTCTCCCTGATAATAACAGATCAGTCCACAGCAATAGCAGCGATTGGCTTTTA

          |         |         |         |         |        420
AAAAGTCCTCCTGCAAATTTATTGGTTCGAAAAGTGATGTTATGCAGGGAAAAGCAGACA

          |         |         |         |         |        480
AGTAATATATGGTCTGTCCTATTTGTAGAAACTTTTTTTACTTACTTGTAGTTTGTGGCC

          |         |         |         |         |        540
ATTACCCATAGATTAAGTTAGTATGTATAGTGGAAGGTAGAAGTATATATATAGTATATA
```

```
        |         |         |         |         |       600
TATCTTGATCTGTCCATGTCTGTCCGTATGAAAGTCGTGTACTAGTCAATGTACATGTAT

        |         |         |         |         |       660
AAAATTTTTAGCTTAACATATCCATCTCTCTCGCGCGGCAATTGGTTGAAGAAACGGGTA

        |         |         |         |         |       720
TCTAGTAGTCGTGACTATAGCGTTCTCCCTTTTTATTTGTTTAATTAATGGGGATAGTCA

        |         |         |         |         |       780
GTATATTGAATTTGACAGAGCATTTAATAGCCAACTTTAGTAAGATTCTAAAATCACCAA

        |         |         |         |         |       840
AATTAGCTGAAACAATAAAATTAAAATTACTATTATTTATTTGAATATTATTACGTATGT

        |         |         |         | intron \ exon C 900
GATCCAAATACATTGACCACTACTGTCCATGACGACTGCCGTCATCAGCCGATATAATGC

        |         |         |         |         |       960
GGAACAGATGTTGCCTGGACCAATTGCAGCCAAATGCCAACGACAATGGCAAATAGCTTG

  Pst I |         |         |         |         |       1020
GAACTGCAGCGCCTAAAAGAAAGAAGTGAAAGGCCAACAAATCCAAACGGAATCGAATGA

        |         |         |         |         |       1080
TCGCCGAGATGCACAAGCAACAGCGCGTGTGGGGAAATATTTTAGAGTGGAAAATCGCTG

        |         |         |         |         |       1140
GCCAAGAATTGAGGAACCGCCGCCAGTCGGCGCCCACTAAACAAACTGAACTTTCACCTC

        |         |         |         |         |       1200
TCAAAACGTAAGACATAGGGATTACCACCTGCCAGGGACTCAGACTCAAGATCACGATGA

        |         |         |         |         |       1260
TGGGGCGACATTGTCATGTGCCAAAGCAAACACACCATATACAATTTGCACAACTTGCCC

     Pvu II     |         |         |         |       1320
AATCCCCAGCTGTCTGTGGATTGTCCGAATCCGCTGCTAATTTACGATACAAGTTATTAT

        |         |         |         |         |       1380
TAATTTGTCATCGGGTCGGCATTAAAATCCCCCGGCAAAAGAAAATACCGACAACTTATA

This splice junction alone fails to obey 'GT-AG' rule
 exon C/ | intron  |         |         |         |       1440
ATGAGCAGGCGTGTATTTGGGACTTTAACATATCCTGTACGATAATGTTTTCAATTCAAG

        |         |         |         |         |       1500
ACATTTACATAAAGCACTCAATTAGGAGAAATAAATGATATATAAAAGGGAATATATTAA

        |         |         |         |         |       1560
TAATTTNATATATATATATATATAAATATACATTAGGTTTTTACTTACAGGTGCCTAAAA

        |         |         |         |         |       1620
GGGGGCACTGAAATAAGGACGTCGTCTTTCAGATGTGATACATTTCTTTTTTCATACTTT

        |         |         |         |         |       1680
TAGCCTACGAGTAGTAATGTTTTTAGCATCCAGATAGACTATGTCTATAGACTATAATGA

        |         |         |         |         |       1740
CTATAAATGACCGTATTGCCGTTTTTATTCGTGTTCCGGGTCAAGATTTTTCATTAAGCG

        |         |         |         |         |       1800
```

AATACACCCCACGATTGG

4.A.2.g Sequence of psBGF (incomplete)

A few bases adjacent to the EcoRI sites at each end of this clone have not been determined

```
          |         |         |         |         |        60
TCTGCCTGTCTGTCTTTTGGATCGCCCGATATTGTGGTAGGTAAACCGTGCGGATAGACG

          |         |         |         |         |       120
TCCCCTAATGGCTCACGGCTAATTGACAAGGGCCAACAAAAATAAATGAAATGGAATGGA

          |         |         |         |         |       180
ATGGTATGGAATTAGCAACGACAATTATGTAGCTGGGGAAGGTCAGTGTTTAATGTCTAT

          |         |         |         |         |       240
GGCTGTATGACATTAATGCTATTTGGCATTGTTCTCGGCTGGAATCGGGGGTCCACAGTC

Xba I Xba I         |        Pvu II     |         |       300
TAGAATCTAGAGTCTAGAATCTAGAGAACAGCTGGCAAATGTCACTTGTCAGGGTATTTT

          |         |         |         |         |       360
TTCGCTTTTACGACAGTTCCAATGCGACATTACGCCCACACGGAACACACCTTTGCAGAT

Bgl II    |         |         |         |         |        \
CTCTATAAATACATATACTACATCCCCACCAATCTCTGCTCTTTATTCGCTTTTTGCAGG

exon D    |         |         |         |         |       480
CAATCTTATTTTATACACCAAGATGGCTGTGGAAATCTTGGGAGGGTGGCAAGATTCATG

                                    A-rich sequence at which
                                    KE2(1.8) was (internally) primed
          |         |         |         | **************540
CGCTCATCATGGACTTAGACATAGGCATTTGTTCCGAAGCCGAGAAAAAACAAAAAAAGA

          |         |    exon D   /    intron     |       600
AATTACTTTTAGATTATTTGTGGGAAAATTTAAGGTAAGTACGGCCCATCCAGCTCTCTG

          |         |         |         |         |       660
ACTAGCCAAATGTACTTAAAGGAGTCCTTTTAACAATAAGTTTTGTTTGAATATTCAAAT

          |         |         |         |         |       720
TTATCCTCCCAATATTGAAGTAATTGAATTTCCAGTTAAATGAAAATTACCAGTAATAGA

          |         |         |         |         |       780
AAAAGTGCAGATAATTTTCCAGTTAAAACCTTAATTACCCGATATATTATTTTCAATAAA

          |         |         |         |         |       840
CACAACCAATGTAAAGCATCGAGTTCAAAAGAATAAGCCTATCTATTAAATGAACCCTTC

          |         |         |         |         |       900
ACAATTTATAATGG
```

4.A.2.h Sequence of psBGJ (incomplete)

```
Eco RI    |         |         |         |         |        60
GAATTCAGGCAGTACAGTGCACTCGAACTCACCTTTGTAGAGGATTTCCAAAACAAACAT
```

```
          |Xho I     |         |         |         |         120
TTCCGGGTAACTCGAGTAGTCGCACTAATTAACAAAGAATTATCCGGAAATCAACTGTGA

          |         |         |         |         |         180
GTCACGAGCCAGAGTTCTTCAGAGACAAAGGATACTCCGCCAAAGGTTTTTGGCATCCAT

          |         |         |         |         |         240
ACATATATTGTATATTTATTTACATCTGAATCACAGCTAATGGCTAAATTTCTGGCGATT

          Eco RV
intron     \        exon E     |         |         |         300
TCGTTTGCAGATATCACAATTGGTGGGCGTACAGATATTACGTGTGTGAGCTGCTCGCCC

      exon E         intron
          |    /    |         |         |         |         360
TTATAAATGTGATAGGTGAGTACTAGTGCACGGCAAACAAAAAAAAAAAAAAACCTAAATA
                                        **************
                                  Poly(A) segment at which SIPC726
                                  has been (internally) primed

          |         |         |         |         |         420
TAATAATAATATCTCAACCGATGGGTCCTGAGATCGTCATGTGCCATAAGGGCCAATCGA

          |         |         |         |         |         480
ACTCAGTTCGCATCATTTATTGCTGCAAG.....remainder of subclone (approximately
1.2 kb) not sequenced
```

4.A.2.i Sequence of psBGK (incomplete)

The sequence of approximately 0.45 kb at proximal end of this subclone has not been determined

```
          |         |         |         |         |         510
TGCACAAATTGGGGGTTAATGGAAAATGAAAGGTGGGAAATCGGTGGCAAAGAACCACTA

          |         |         |         |         |         570
TCAATCCCTGTGTCCGGTCCGATAAACAGATAGGCAAGTAGGACAGATGAATAGGCAGAT

          |         |         |         |         |         630
AAACCGAACGAACGAAAAGNGTATCTCTGTATCTCTTTGGTGCTGATTTGTATCTTTTTT

      intron  \  exon F         |         |         |         690
GTGGTTCCTTTGCAGGTCAAATGTTTCTTATGAATCGATTTTTCGATGGCGAATTTATAA

          |         |         |         |         |Bam HI 750
CATTTGGCCTGAAAGTGATAGATTATATGGAGACCGATCAGGAAGATCGCATGGATCCGA

                                                exon F
          |         |         |         |         |         810
TGATTTACATATTTCCCAGAATGACCAAATGTACATTTTTTAAATATGGTTCCAGTGGGG

       intron
 /        |         |         |         |         |         870
AGGTGAGTAACGAATGCAGCACTCGACAAACCCAAGACTCAATAATAATTTAATCTTTCT

  \ exon G          |         |         |         |         930
AGGTGGAGAAACACGACGCCATTTGCATTTTACCATTAAACGTTGTTAATGAGAAGATTT

          |         |         |         |         |         990
```

```
ACATTTTCCTTTGGTTTTGGTTTATATTATTAACGTTTCTCACATTATTAACGCTAATAT

        |         |         |     Acc I          exon G  1050
ACAGGGTGGTTATTATATTCTCTCCTCGAATGAGGGTCTACTTATTTCGTATGCGATTTA

      intron
/       |         |         |         |         |       1110
GGTGAGTTTCATAGACATTGCGGGTCCTCACTGGCGGTGATATGAGAACTAAAAAAGGGA

      Hin dIII  |         |         |         |         1170
AAATGTAAAGCTTATTAAAAAGTTGCTTTCGAGCTAAGGATTTGCGTTTCTATGAGATGT

        |         |         |         |         |       1230
GACTTATGTATGACTTATGTATGTGGGTTCCACAAAAATCCCTATACTCGTATAGGTAGG

        |         |         |         |         |       1290
TGCCTAGAAGTATGTAGCCAAAACAATTTTTTTTTCACTACATACTTCGTAGAGCCCTAT

        |         |         |         |         |       1350
GTGAAATTTAAGCCTTAGATTTTTCCGTGACACCGCCCAGGTTTCTTGTCAATATCTGTT

        |         |         |         |         |       1410
TCCCTATCAAGCTAGTGATTCCCATTAGCTTACATAATGATTTGAACAAACTTACTTAAT

    intron          exon H
        |         \         |         |         |       1470
GCCATTTCTTTAATTGCAGGTTAGTGCGTCGTGACGCTATTGAAATAATCGTTCGTCGTT

        |         |         |         |         |       1530
CAAAGATGGGCGATTGGTTTTTGTTGTATTTACTAGGTGAAAACATAGATACAGTTATAT

        |         |         |         |         |       1590
TTCGTGATGTTGTACAGGACTTAGCGAATCGTTTAGGACATAACCAACACCACAGGGTGC

        |         |         |         |         |       1650
CTGGCTTAAAAGGTGAAATACAGGATGCATGATATTGGGAGTATTAGAAACAATACAAAA

        |         |         | Eco RI  |         |       1710
TGCAATTTGTCGTCTCCATTTAAAAACCATCGAATTC
```

4.A.2.j psBGM sequence (incomplete)

```
Same Eco RI site here as at end of psBGK; exon H continues across
the boundary between subclones K and M
Eco RI    |         |         |         |         |        60
GAATTCAATAACAAAAATGTCGAAAGCAAGAAAAAGATTAAGAAGGACAATTACAACCAC

        |         |         |         |         |       120
AAAGGAATCTAGAGATCTTCGCAGCAGCCGCTTCATTAAAACTTACAACTCAACACACCG

poly(A) segment in which SIPC8, SIPC737 and B10 were internally primed
        |   **********      |         |         |       180
CTAAAAAATCTTTAAAAAAAAAACGTTATTCAAAATCACCAAAAACCT.....
```

Remainder of subclone (approximately 2.4 kb), not sequenced.

4.A.3 The M23 transcription unit

The M23 transcription unit is derived from a region some 20 to 35 kilobases proximal to the start of KE2(1.8), proximal to the proximal breakpoint of *Df(1)16-3-35* (Figure 3.2; §4.6). M23 also is very likely to be transcribed in the opposite (distal to proximal) direction to *shaking-B*. As mentioned above (§4.6) the M23 transcription unit was studied before the true identity of the *shaking-B* gene became clear. Subsequent to the identification of *shaking-B*, little further work has been done on M23. For this reason the analysis of M23 is incomplete.

4.A.3.a The M23 cDNA family

M23 itself is the largest of a family of overlapping cDNA clones which were isolated from a *Drosophila* adult cDNA library (Stratagene) by Helena Yang (M11B, M12A, M13, M14, M23) and Shuqing Ji (S3A). These cDNAs differ only in their lengths, no differences in splicing patterns being evident. When the M23 cDNA is hybridised to restriction enzyme digested phage clones from the 952 chromosome walk, hybridisation is non-contiguous (Helen Clinkenbeard, unpublished results; not shown), suggesting that M23 and its siblings contain introns and are therefore *bona fide* cDNAs. Restriction maps of M23 and its smaller relatives are shown in Figure 4.A.1.a.

Figure 4.A.1.a: The M23 cDNA family. See next page for legend.

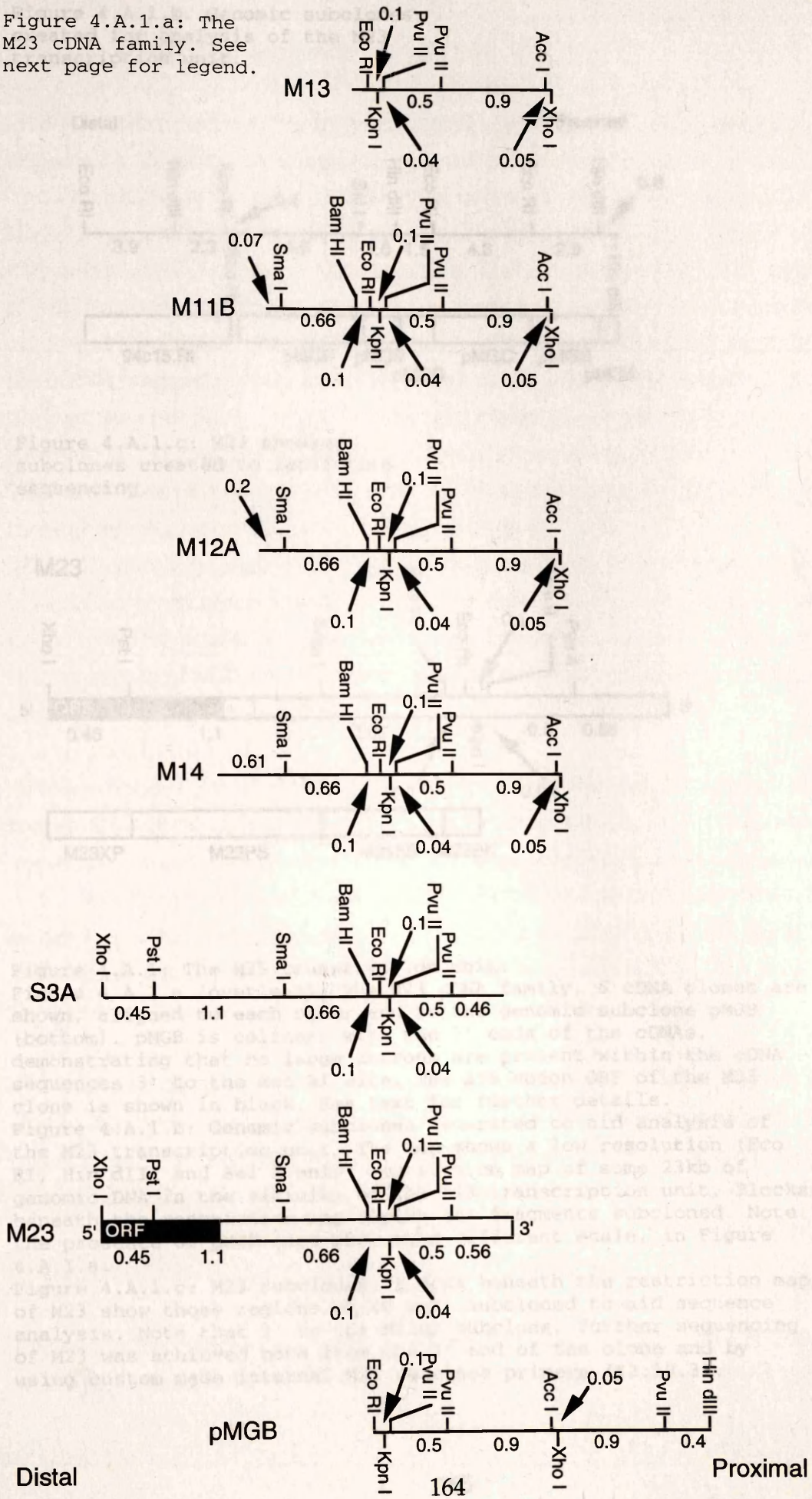


Figure 4.A.1.b. Genomic subclones created for analysis of the M23 transcription unit.

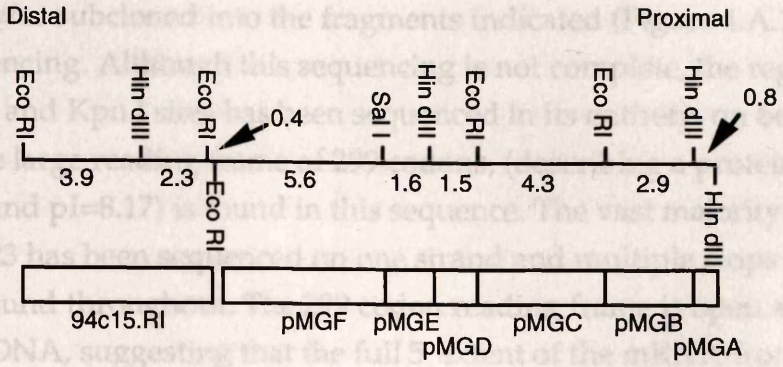


Figure 4.A.1.c: M23 showing subclones created to facilitate sequencing.

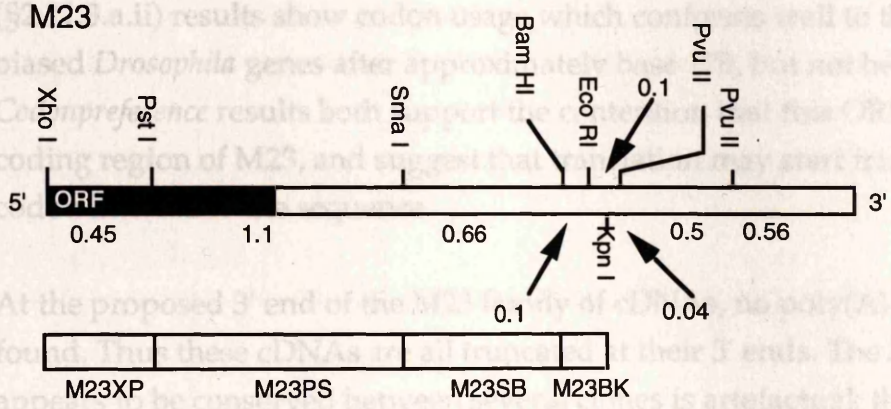


Figure 4.A.1: The M23 transcription unit.
Figure 4.A.1.a (overleaf): The M23 cDNA family. 6 cDNA clones are shown, aligned to each other and to the genomic subclone pMGB (bottom). pMGB is colinear with the 3' ends of the cDNAs, demonstrating that no large introns are present within the cDNA sequences 3' to the Eco RI site. The 299 codon ORF of the M23 clone is shown in black. See text for further details.
Figure 4.A.1.b: Genomic subclones generated to aid analysis of the M23 transcription unit. The map shows a low resolution (Eco RI, Hin dIII and Sal I only) restriction map of some 23kb of genomic DNA in the vicinity of the M23 transcription unit. Blocks beneath the restriction map depict the fragments subcloned. Note the presence of pMGB here and, at a different scale, in Figure 4.A.1.a.
Figure 4.A.1.c: M23 subclones. Blocks beneath the restriction map of M23 show those regions which were subcloned to aid sequence analysis. Note that 3' to the M23BK subclone, further sequencing of M23 was achieved both from the 3' end of the clone and by using custom made internal M23 sequence primers (§2.10.3).

4.A.3.b The M23 sequence

M23 was subcloned into the fragments indicated (Figure 4.A.1.c), to facilitate its sequencing. Although this sequencing is not complete, the region between the Xho I and Kpn I sites has been sequenced in its entirety, on both strands. A single large reading frame of 299 codons, (describing a protein fragment of 33.6 kDa and pI=8.17) is found in this sequence. The vast majority of the remainder of M23 has been sequenced on one strand and multiple stops in all three frames are found throughout. The 299 codon reading frame is open at the very start of the cDNA, suggesting that the full 5' extent of the mRNA from which M23 was derived may not be represented in the M23 clone. However, the *Testcode* program (§2.11.3.a.i; results not shown) shows low scores at the start of this ORF until base positions 150-200, after which scores climb and remain high throughout the rest of the reading frame. Similarly, *Codonpreference* (§2.11.3.a.ii) results show codon usage which conforms well to that of highly biased *Drosophila* genes after approximately base 170, but not before. These *Codonpreference* results both support the contention that this ORF is indeed the coding region of M23, and suggest that translation may start from an initiation codon internal to the sequence.

At the proposed 3' end of the M23 family of cDNAs, no poly(A) regions are found. Thus these cDNAs are all truncated at their 3' ends. The 3' end which appears to be conserved between several clones is artefactual: the construction of the Stratagene *Drosophila* adult cDNA library relied upon hemimethylation of newly synthesised cDNA strands to protect internal sites against digestion with Xho I. The common 3' end observed is at the position of an Xho I site in the genome, demonstrating that this protection failed.

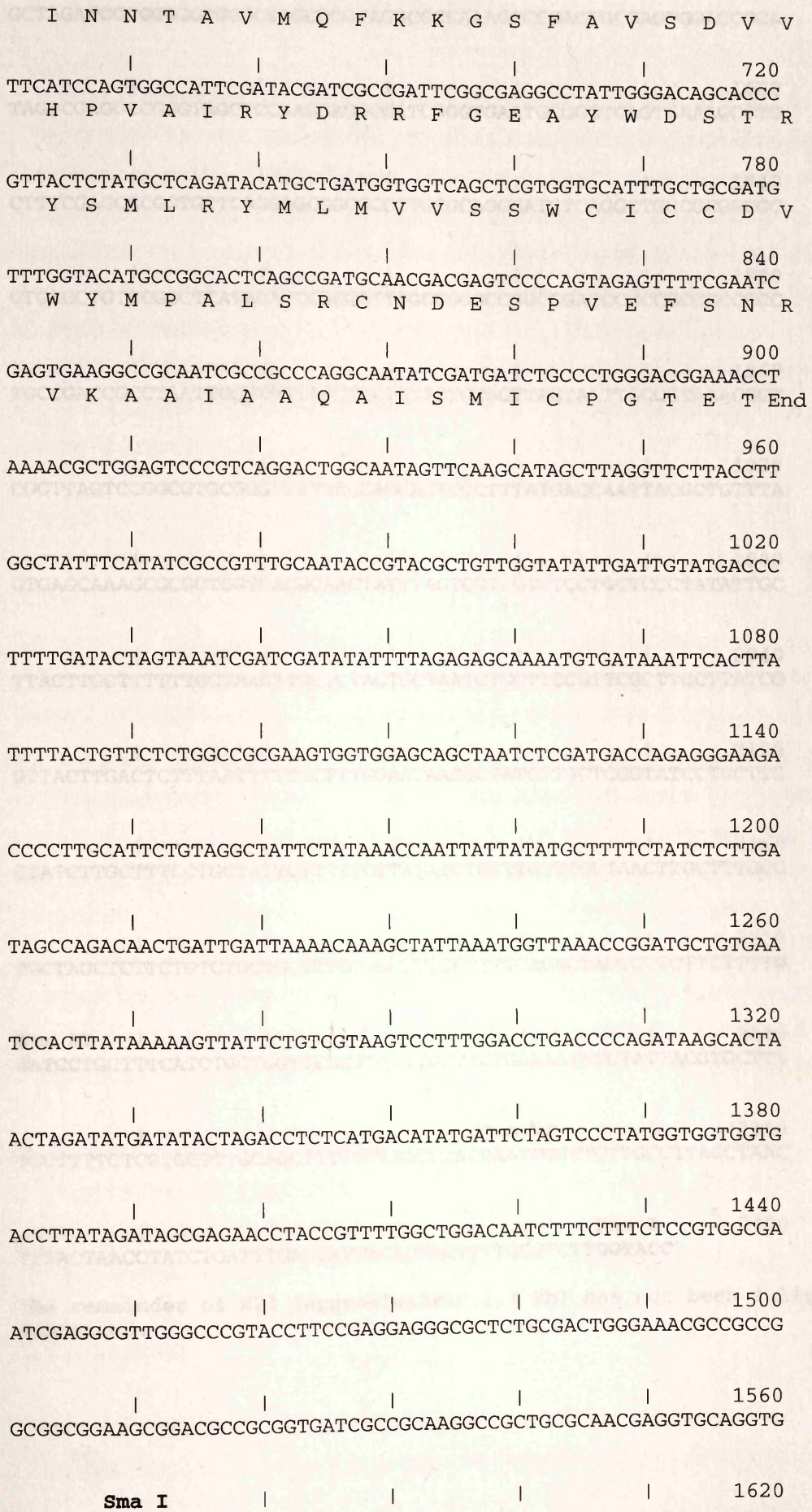
The M23 and S3A clones share the same apparent 5' end. This may either be due to the presence of another genomic Xho I site (the relevant genomic region has not yet been mapped) or may reflect the genuine 5' extent of transcripts from which these clones were derived. If this latter possibility is indeed the case, then translation must initiate from a site internal to the existing sequence, as suggested by *Testcode* and *Codonpreference* (see above).

Regarding the protein structure implied by the M23 sequence, two points are worthy of mention. Firstly, the M23 transcription unit does not show significant

homology to any entries in the sequence databases tested (§2.11.3.b), either at the DNA or at the protein level. Secondly, the 299 codon ORF of M23 may encode a transmembrane protein. All transmembrane prediction algorithms used (§5.1.2) predict at least one transmembrane domain, at around residues 57-77 in the putative M23 peptide sequence, while the Engleman (§5.1.2.b) and von Heijne (§5.1.2.e) techniques predict a further transmembrane helix at residues 246 to 266. More detailed analysis of M23 protein structure is of limited value until it is clear whether or not the 5' end of M23 reflects that of the transcript from which the cDNA was derived.

The (incomplete) sequence of M23 is presented below:

```
Xho I     |         |         |         |         |        60
CTCGAGGACGACGTGACCCAGAGGTTTGTGGCCCCACCTTCGCCAGCTGGGGAATGGAAT
   R  G  R  R  D  P  E  V  C  G  P  T  F  A  S  W  G  M  E  S
   M23(33.6) ORF

          |         |         |         |         |       120
CTGCTGACCCGCAACCTTCGTCAGAGGAATCGATACCTCAGCTGGCGTTTAAGGACGGTC
   A  D  P  Q  P  S  S  E  E  S  I  P  Q  L  A  F  K  D  G  L

          |         |         |         |         |       180
TGGCTGCTGGGCTGGGTTGTACGCTACGGACTGCTTTTGCCCTTCAGGACCATTGGCTGC
   A  A  G  L  G  C  T  L  R  T  A  F  A  L  Q  D  H  W  L  L

          |         |         |         |         |       240
TGGCTGTGCTGTTCATGATCAGTGGAGTCTCAATGCTGCTGGGTCACATTCCCGACTGGT
   A  V  L  F  M  I  S  G  V  S  M  L  L  G  H  I  P  D  W  C

          |         |         |         |         |       300
GCTTCAAGAAAAAGTTGGTGGAGCTGGTATTGCGGCAGTGCTTCCGAATCACGGCCGCCT
   F  K  K  K  L  V  E  L  V  L  R  Q  C  F  R  I  T  A  A  C

          |         |         |         |         |       360
GCCTGCCGATGATCCGCAGATTCCACAACACGGAGTATCGGCCCACCAAGGGCATCTGTG
   L  P  M  I  R  R  F  H  N  T  E  Y  R  P  T  K  G  I  C  V

          |         |         |         |         |       420
TGTGCAATCACACGAGCCCCTTAGACGTCCTGGTGCTGATGTGCGATGCCAACTACTCGC
   C  N  H  T  S  P  L  D  V  L  V  L  M  C  D  A  N  Y  S  L

          |         |         |    Pst I          |       480
TGACGGGCCAGGTTCACACCGGCATCCTGGGCGTCCTGCAGCGCGCCCTGTCGAGGGTCT
   T  G  Q  V  H  T  G  I  L  G  V  L  Q  R  A  L  S  R  V  S

          |         |         |         |         |       540
CCCACCACATGTGGTTCGACCGCAAGGAACTGGCCGATCGTGAGGCTCTGGGTCTGGTGC
   H  H  M  W  F  D  R  K  E  L  A  D  R  E  A  L  G  L  V  L

          |         |         |         |         |       600
TCCGCCTACACTGCTCCATGAAGGATCGGCCGCCGGTCCTGCTCTTCCCCGAGGGCACTT
   R  L  H  C  S  M  K  D  R  P  P  V  L  L  F  P  E  G  T  C

          |         |         |         |         |       660
GCATCAACAACACCGCCGTAATGCAGTTCAAGAAGGGTAGCTTCGCGGTCAGCGACGTCG
```


I N N T A V M Q F K K G S F A V S D V V
720
TTCATCCAGTGGCCATTCGATACGATCGCCGATTCGGCGAGGCCTATTGGGACAGCACCC
H P V A I R Y D R R F G E A Y W D S T R
780
GTTACTCTATGCTCAGATACATGCTGATGGTGGTCAGCTCGTGGTGCATTTGCTGCGATG
Y S M L R Y M L M V V S S W C I C C D V
840
TTTGGTACATGCCGGCACTCAGCCGATGCAACGACGAGTCCCCAGTAGAGTTTTCGAATC
W Y M P A L S R C N D E S P V E F S N R
900
GAGTGAAGGCCGCAATCGCCGCCCAGGCAATATCGATGATCTGCCCTGGGACGGAAACCT
V K A A I A A Q A I S M I C P G T E T End
960
AAAACGCTGGAGTCCCGTCAGGACTGGCAATAGTTCAAGCATAGCTTAGGTTCTTACCTT
1020
GGCTATTTCATATCGCCGTTTGCAATACCGTACGCTGTTGGTATATTGATTGTATGACCC
1080
TTTTGATACTAGTAAATCGATCGATATATTTTAGAGAGCAAAATGTGATAAATTCACTTA
1140
TTTTACTGTTCTCTGGCCGCGAAGTGGTGGAGCAGCTAATCTCGATGACCAGAGGGAAGA
1200
CCCCTTGCATTCTGTAGGCTATTCTATAAACCAATTATTATATGCTTTTCTATCTCTTGA
1260
TAGCCAGACAACTGATTGATTAAAACAAAGCTATTAAATGGTTAAACCGGATGCTGTGAA
1320
TCCACTTATAAAAAGTTATTCTGTCGTAAGTCCTTTGGACCTGACCCCAGATAAGCACTA
1380
ACTAGATATGATATACTAGACCTCTCATGACATATGATTCTAGTCCCTATGGTGGTGGTG
1440
ACCTTATAGATAGCGAGAACCTACCGTTTTGGCTGGACAATCTTTCTTTCTCCGTGGCGA
1500
ATCGAGGCGTTGGGCCCGTACCTTCCGAGGAGGGCGCTCTGCGACTGGGAAACGCCGCCG
1560
GCGGCGGAAGCGGACGCCGCGGTGATCGCCGCAAGGCCGCTGCGCAACGAGGTGCAGGTG
Sma I
1620

GCTAGAT**CCCGGG**GGTGGGCTAGCCCGTAGGCGCCAAAGGCCGACTGCGAGTGGACCGCA

1680
TAGTCGAGGCCGGGTAGCTCCAAGGAGGCGATCGGGTGAATGCGGGTCAGTGAAAGCGTG

1740
CTTTCGAGCACGGTGCTCAGGAGCGGGGCGCTGCGCAGGGATCTCAGGCTGCCGCGGCGC

1800
GTGAGCTGTCCGGCTTATACATCCGGGGTTGGGGGGGCCGGCCGACCCCCCGCCGCCGCC

1860
TGCCGACCGCCTAATTGCTTGTCACCTCCTCTGCTCCGGTTAGTAATTGCCGTAAACGGT

1920
CGGTTAGTCCGGCGTGCGGGTCATTCGCAGCAGTCGCTTTATGACCAATTACGCTGTTTA

1980
GTGAGCAAAGCGCGCTGGTCACGCAACTATTTAGTCGTCGTCTCCTGCTCCCTATATTGC

2040
TTACTTGCTTTTTTGCTAACTTGCTCTACTGCTAATCTGCTTGCGTTCGCTTCCTTATCG

2100
GTTACTTGACTCTTTAATTTTTGCTTTGGAATAAGGCTATCTTGCTCGGTATCCTGCTTG

2160
CTATCTTGCTTTCCTGCTATTCTTTTTGTTATATCTGCTTGCTTGCTAACTTGCTTTGCC

2220
TGCTAGCTCTTCTGTCTGCTGCTTTCTTGCTTTCCTCTTCAGACTAGTCTTCTTCTTTT**G**

Bam HI 2280
GATCCTGCTTTCATCTGCTGGTTCGCTTTCTTGCTTCTGAAAATATCTATTACGTGCTTT

Eco RI 2340
TGCTTTTCTCGTGCTTTGCAGCTTTTGTTAGCCTAC**GAATTC**CTTCTTGCCTTACCTAAC

Kpn I 2400
TTTACTAACCTATCTGATTTGACTATTACACTACTTTTGCGTCTT**GGTACC**

The remainder of M23 (approximately 1.1 kb) has not been fully sequenced.

Five

The results presented in chapters 3 and 4, together with those of Krishnan and coworkers (Krishnan, *et al.*, 1993) provide a substantial body of data regarding the structure of the *shaking-B* gene. The strong homologies identified among the Shaking-B, Ogre and Unc-7 proteins are tantalising and yet the major hurdle of elucidating the biochemical functions of Shaking-B and its relatives still looms large. To help negotiate this hurdle, this chapter is devoted to a detailed analysis of the structures of Shak-B proteins and their homologues. The analysis will focus on two main aspects: the prediction of potential membrane spanning segments and their topologies (§5.1-§5.4) and the search for peptide motifs of known function (§5.5).

5.1 ARE ANY SHAK-B PROTEINS SECRETED OR MEMBRANE BOUND ?

Many of the molecules implicated to date in the establishment of synapses are either secreted (e.g. Matthes, *et al.*, 1995) or membrane bound (e.g. Chiba, *et al.*, 1995) and it is therefore of particular interest to discover whether some or all of the products of *shaking-B* are themselves secreted or membrane associated. In eukaryotes, most secreted proteins are translated as preproteins bearing an N terminal signal sequence (for reviews see Gierasch, 1989; von Heijne, 1990; for rare exceptions see Muesch, *et al.*, 1990). On translocation of the secretory preprotein through the endoplasmic reticulum (a process which may be either cotranslational or posttranslational depending upon the protein concerned (Jungnickel, *et al.*, 1994)), the signal peptide is removed by signal peptidase, leaving the mature protein in the secretory compartment. For recent reviews of the protein translocation process, see Jungnickel, *et al.*, 1994; Rapoport, 1992.

Membrane spanning proteins may be classified into four groups (classification of (von Heijne and Gavel, 1988)) according to the positions of their N and C termini, and the number of times they span the membrane (see Figure 5.1). Class I, II and III proteins span the membrane once only and are sometimes referred to as *bitopic* proteins. Class IV (*polytopic*) proteins span the membrane multiple times.

Type I proteins are translated as preproteins with cleaved N terminal signal peptides. Their membrane translocation is wholly analogous to that of most

secreted proteins except that an internal, hydrophobic, membrane spanning "stop transfer" sequence arrests their translocation (Rapoport and Wiedmann, 1985), leaving the protein anchored in the membrane with its N terminus extracytoplasmic and its C terminus cytoplasmic. This is by far the commonest class of bitopic membrane proteins.

Type II proteins have uncleaved signal sequences by which they are anchored into the membrane, hence the name *signal anchor* sequences (Lipp and Dobberstein, 1988). By definition, their orientation is opposite to that of type I proteins, their N termini being cytoplasmic. In addition to being uncleaved, signal sequences of type II proteins differ from those of secreted proteins and type I molecules in two respects, both of which reflect their rôle as transmembrane domains. Firstly type II signal sequences generally contain longer hydrophobic stretches than their type I counterparts (Hegner, *et al.*, 1992; von Heijne, 1985). Secondly, type II signals may reside anywhere within the protein, though they are usually closer to the N than to the C terminus.

Type III proteins are rare. For a catalogue of the majority of those currently known, see (Parks and Lamb, 1993). Like type I proteins they have extracytoplasmic N termini and cytoplasmic C termini, but unlike type I proteins, they acquire this topology in the absence of a labile N terminal signal peptide. Type III proteins, like type II, are directed to the endoplasmic reticulum membrane by a signal anchor sequence which may reside anywhere within the protein, but in this class of proteins the signal anchor is often close to the C terminus.

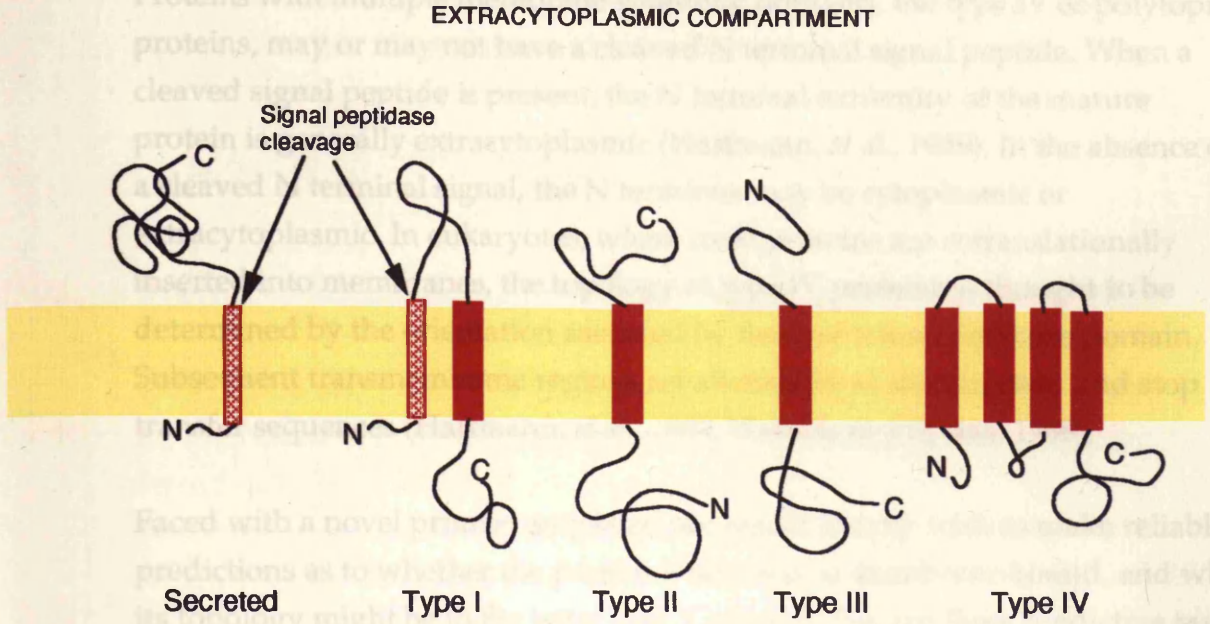


Figure 5.1: Secreted and integral membrane proteins. Classification is that of von Heijne and Gavel (1988). Hatched red rectangles represent cleaved N-terminal signal peptides. Filled red rectangles represent transmembrane domains. While most eukaryotic secreted proteins have the general structure represented here, exceptions do exist (see Muesch *et al.*, (1990) for a review). Type IV proteins are any that span the membrane multiple times. Only one representative of this class is shown.

Proteins with multiple membrane-spanning domains, the type IV or polytopic proteins, may or may not have a cleaved N terminal signal peptide. When a cleaved signal peptide is present, the N terminal extremity of the mature protein is generally extracytoplasmic (Hartmann, *et al.*, 1989). In the absence of a cleaved N terminal signal, the N terminus may be cytoplasmic or extracytoplasmic. In eukaryotes, where most proteins are cotranslationally inserted into membranes, the topology of type IV proteins is thought to be determined by the orientation assumed by the first transmembrane domain. Subsequent transmembrane regions act alternately as start transfer and stop transfer sequences (Hartmann, *et al.*, 1989; Wessels and Spiess, 1988).

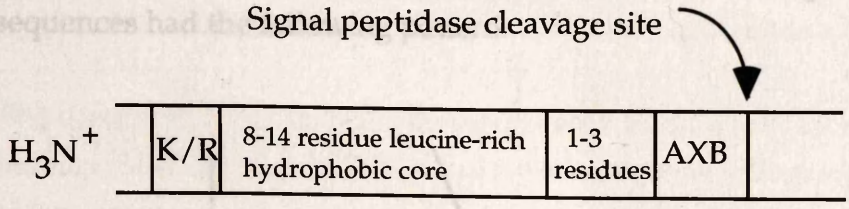
Faced with a novel primary sequence, we would ideally wish to make reliable predictions as to whether the protein is secreted or membrane bound, and what its topology might be in the latter case. Central to this are three predictive tasks:

(a) Identification of potential N-terminal cleaved signal sequences
(b) Prediction of transmembrane domains, including signal anchors
(c) Prediction of the orientations of any transmembrane segments

These three key predictions will be considered in some depth and results obtained for Shaking-B proteins and their homologues will be discussed.

5.1.1 Searching for cleaved N-terminal signal peptides

Recombinant proteins comprising a cleavable signal peptide from one organism and a mature secretory protein from another are frequently export competent (e.g. Mueller, *et al.*, 1982), demonstrating that at least some signal peptides must have common biochemical features. Despite this, however, signal peptides display a striking lack of primary sequence homology, even among closely related proteins (Gierasch, 1989), thus complicating the identification of signal peptides within newly derived gene sequences. Several methods of cleaved N terminal signal peptide prediction have been proposed (McGeoch, 1985; Perlman and Halvorson, 1983; von Heijne, 1986). The seminal work in this field is that of Perlman and Halvorson. These authors analysed the signal sequences of 39 proteins, and proposed a number of common structural motifs: an N-terminal region of 11 residues or less, preceeding the signal, a basic (R or K) residue, followed by a run of 8-14 hydrophobic residues, then a linker region of 1-3 residues before the cleavage site:



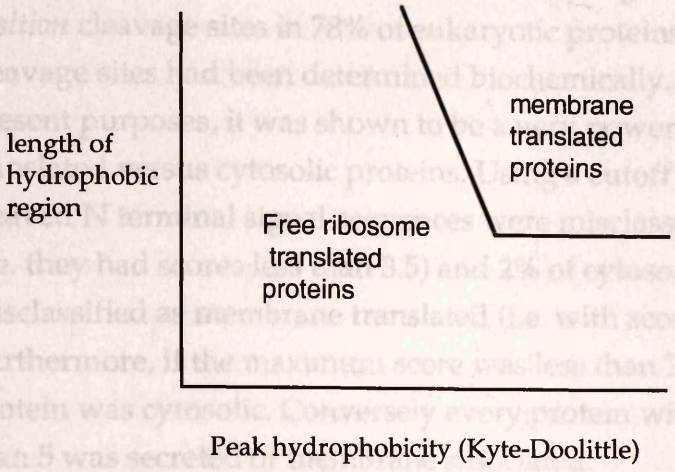
where A= Ala, Gly, Ser, Leu or Val; B= Ala, Gly, or Ser; X=any residue.

Using Chou and Fasman structural predictions (Chou and Fasman, 1978), these authors found that the hydrophobic core region tends to have an extended β strand as its most favoured structure, though α helices are also predicted to be possible and in a non polar environment within the membrane, α helical structures may be strongly favoured (see Rosenblatt, *et al.*, 1980). They also observed a strong tendency for a β turn close to the signal petidase cleavage site, and argued that this supported the "reverse hairpin" or "loop" model of signal peptide structure proposed by other laboratories (Engelman and Steitz, 1981; Shaw, *et al.*, 1988).

The work of Perlman and Halvorson provided the first confident assertions of structural motifs common among cleaved N terminal signal peptides. In terms, however, of providing a predictive method for the recognition of signal peptides in newly derived sequences, the usefulness of their findings was limited, as the investigator was left with a number of structural features to look for, and the extent of conformity required before a sequence may be identified as a cleaved N terminal signal sequence was an open question demanding a subjective answer. Moreover, only 39 signal peptide sequences were available to Perlman and Halvorson, thus all of their predictions were made on the basis of a very low sample size.

The empirical method of McGeoch (1985) was developed using 114 cleaved N teminal signal sequences and provides a less subjective analytical tool for their prediction. The N terminus of a query protein was first classified into charged and uncharged regions by assigning the residue following the last charged residue of positions 1 to 11 as the first residue of the uncharged region, and calling it position 12. Starting from position 12, the length of the uncharged region was defined as the number of residues before a K, R, D, E, or H was encountered. This statistic was plotted against the Kyte-Doolittle hydrophobicity (Kyte and Doolittle, 1982) of the most hydrophobic 8 residue

window in positions 12 to 21. The trend obtained for a large number of query sequences had the following pattern:



The dividing line between membrane translated and free ribosome translated products was an arbitrary one found to effectively separate these two classes of proteins. The predicitive value of McGeoch plots is fairly high. If a sequence of interest falls comfortably into the membrane bound region then it is quite probably a cleaved N terminal signal sequence. Thus the *selectivity* of the predicition is high. Unfortunately some *bona fide* signal sequences fall outwith the membrane translated region on McGeoch plots, thus the *sensitivity* of the search tool is perhaps too low. McGeoch did not address the prediction of uncleaved N terminal signal anchors, few if any such sequences having been demonstrated by 1985.

An entirely different approach to the signal sequence recognition problem was taken by von Heijne (1986). His approach was based on weight matrices (Staden, 1984; §3.5.b.iii). Von Heijne analysed 161 eukaryotic, cleaved N terminal signal peptides and aligned them according to their signal peptidase cleavage sites to give a table of amino acid counts in each position $N(a,i)$ (i.e. the number of residues of type a at position i). Weight matrices $W(a,i)$ were calculated by taking the observed amino acid counts in each postition and dividing them by their expected abundance in proteins in general (Klapper, 1977), then taking the natural logarithm of these quotients. Because there is no logarithm of zero, adjustments had to be made to allow for zero entries in the abundance matrix. The predictive value of residues -13 to +2 relative to the cleavage site was found to be greatest, and query sequences were scanned with

this 15 residue window. The highest scoring window within the N-terminal 40 residues is taken to be a signal peptide, if its score exceeds a threshold value.

The accuracy of von Heijne's method is very high. It was found to correctly *position* cleavage sites in 78% of eukaryotic proteins whose signal peptidase cleavage sites had been determined biochemically. More importantly for present purposes, it was shown to be a very powerful predictor of membrane translated versus cytosolic proteins. Using a cutoff score of 3.5, 2% of *bona fide* cleaved N terminal signal sequences were misclassified as cytosolic N termini[1] (i.e. they had scores less than 3.5) and 2% of cytosolic proteins were misclassified as membrane translated (i.e. with scores greater than 3.5). Furthermore, if the maximum score was less than 2 then, in all cases tested, the protein was cytosolic. Conversely every protein with a maximum score greater than 5 was secreted or membrane associated.

5.1.2 The *Signify* program

The advantages of von Heijne's method are that the analysis throws out a single statistic for each query sequence, and this gives an answer of cytosolic, membrane translated, or unable to classify. The only major disadvantage is that the method requires a lot of calculation which is not feasible to do by hand. After many unsuccessful attempts to track down a computer program based on von Heijne's technique, I undertook to write one. The result was the *Signify* 1.1 program, written for Macintosh in the THINK-C environment (Symantec) to detect eukaryotic N terminal cleaved signal sequences in query proteins. The *Signify* code is presented in complete and annotated form in the appendix to this chapter. This double clickable application will run on any Macintosh running version 7.0 or greater of the System software and uses 384K of RAM. Since completing *Signify* in a spartan though highly functional form, I have discovered two related programs, *Signalase*, by Ned Mantei, designed to search for *mammalian* signal peptides, and *Plota_Sig*, part of the *Plota* protein analysis package of A. Luettke and P. Markiewicz. The former can be obtained from the University of Indiana IUBIO archive, while the latter is available via the EMBL file server (Fuchs, *et al.*, 1990). *Signalase* is restricted to the analysis of

---

1 The estimation of the number of membrane translated sequences misclassified by this method was based on testing 161 membrane-translated N termini, of which three were misclassified. There is, however, an intrinsic bias in this estimate, as the same 161 sequences were used to construct the frequency table. Even given this, the false negative rate is likely to be less than 3%, a remarkable achievement. This bias does not, of course apply to the false positive rate.

mammalian signal peptides, and is therefore not appropriate here. *Plota_Sig*, is a very user friendly implementation of the von Heijne method. *Plota_Sig*, will propose the same 'best signal peptides' as *Signify* (as would be expected) but I have found the scores produced by *Plota_Sig*, to be much higher than the von Heijne algorithm should generate, despite the claim that the same scoring scale is used. I have therefore used *Signify* in all of the signal peptide predictions in this work.

Figure 5.2 shows a flow diagram of the *Signify* program. *Signify* scans through a peptide sequence read from a disk file or input from the keyboard, and prints out every sequence window with a signal peptide prediction score greater than a threshold which is input by the user. The program does not restrict itself to scanning only the N terminal 40 residues of an input sequence, to allow investigation of how internal signal anchor sequences are classified (see discussion below). The format of sequences in disk files is not critical as long as the sequence itself is the only text in the file, however text only, upper case sequences, without spaces or numbers are interpreted most quickly. *Signify* is only able to look for files in the same folder as the application, so query sequences should be copied into the *Signify* folder before the program is launched. Furthermore, due to the nature of the string in which the protein filename is input, *Signify* will only interpret filenames which do not contain non-alphabet characters. Once the program is completed, a hard copy is output to the printer specified in the Chooser. This copy includes a simple dot histogram which makes it easy to detect the highest scoring window in the query protein.

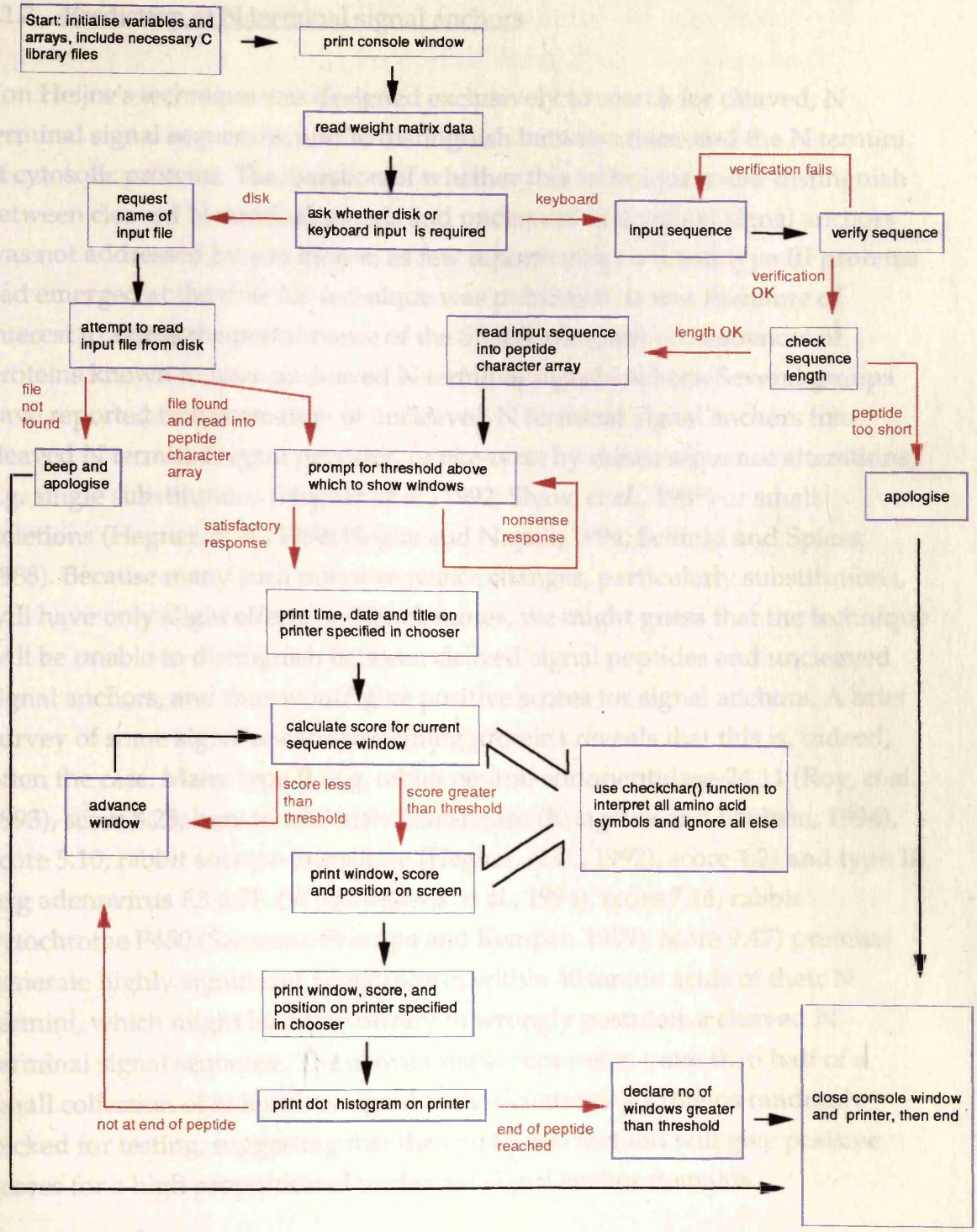


Figure 5.2: Flow diagram of the *Signify* program. For a full program listing of the *Signify* code, see the Appendix to this chapter.

5.1.3 Prediction of N terminal signal anchors

Von Heijne's technique was designed exclusively to search for cleaved, N terminal signal sequences, and to distinguish between these and the N termini of cytosolic proteins. The question of whether this technique could distinguish between cleaved N terminal signals and uncleaved N terminal signal anchors was not addressed by von Heijne, as few reports of type II and type III proteins had emerged at the time his technique was published. It was therefore of interest to assess the performance of the *Signify* program on sequences of proteins known to have uncleaved N terminal signal anchors. Several groups have reported transformation of uncleaved N terminal signal anchors into cleaved N terminal signal peptides, or *vice-versa* by minor sequence alterations, e.g. single substitutions (Hegner, *et al.*, 1992; Shaw, *et al.*, 1988) or small deletions (Hegner, *et al.*, 1992; Hogue and Nayak, 1994; Schmid and Spiess, 1988). Because many such minor sequence changes, particularly substitutions, will have only slight effects on *Signify* scores, we might guess that the technique will be unable to distinguish between cleaved signal peptides and uncleaved signal anchors, and thus would give positive scores for signal anchors. A brief survey of some signal anchor-containing proteins reveals that this is, indeed, often the case. Many type II (e.g. rabbit neutral endopeptidase-24.11 (Roy, *et al.*, 1993), score 8.25; human α2,3-sialyltransferase (Kitagawa and Paulson, 1994), score 5.10; rabbit sucrase-isomaltase (Hegner, *et al.*, 1992), score 4.2) and type III (e.g adenovirus E3-6.7K (Wilson-Rawls, *et al.*, 1994), score 7.16; rabbit cytochrome P450 (Szczesna-Skorupa and Kemper, 1989), score 9.47) proteins generate highly significant *Signify* scores within 40 amino acids of their N termini, which might lead the unwary to wrongly postulate a cleaved N terminal signal sequence. The above sample represents more than half of a small collection of N terminal signal anchor-containing proteins randomly picked for testing, suggesting that the von Heijne method will give positive scores for a high proportion of uncleaved signal anchor domains.

There is no evidence to suggest that this detection by *Signify* of signal anchor domains reflects any compositional bias in signal anchors as compared to other transmembrane domains. Compositional biases of signal anchors compared to type I transmembrane domains have been sought but not found (von Heijne and Gavel, 1988), and the only critical sequence requirement for signal anchor function is identical to that of other α helical transmembrane domains: a stretch of amino acid residues long enough and hydrophobic enough to span a

membrane (Zerial, *et al.*, 1987) see §5.1.2. The corollary of this: that transmembrane domains other than signal anchors may also generate high *Signify* scores, is often the case, but in practice this causes little confusion as it is unusual for a transmembrane domain which is not a signal anchor to be present in the first 40 residues of a polypeptide.

Because there are no detectable differences in compositional bias between signal anchors and type I transmembrane domains (von Heijne and Gavel, 1988), techniques dedicated to the identification of transmembrane regions (§5.1.2) may be used for the prediction of signal anchor sequences. However confusion is generated by the fact that cleaved N-terminal signal sequences are often positively identified by transmembrane domain prediction algorithms. In one study (Hartmann, *et al.*, 1989) 149 out of 200 cleaved signal peptides were identified as transmembrane regions by the prediction methods used. However, cleaved N terminal signal sequences tend to contain shorter hydrophobic stretches (averaging 8-12 hydrophobic residues compared to around 20 in signal anchors (Hegner, *et al.*, 1992)). They therefore tend, when identified by transmembrane helix prediction methods, to have borderline scores. This is considered further below (§5.4.2).

5.1.4 Prediction of eukaryotic signal peptides: A summary

The von Heijne prediction technique (von Heijne, 1986), as implemented in the *Signify* program is the method of choice for the detection of N terminal signal sequences, but does not discriminate effectively between cleaved signal peptides and uncleaved signal anchors. Because signal anchors are indistinguishable in sequence characteristics from other transmembrane α helices (von Heijne and Gavel, 1988), they are identifiable with techniques dedicated to transmembrane domain prediction (§5.1.2), though cleaved signal peptides often generate false positives using these techniques (Hartmann, *et al.*, 1989). If a query sequence contains within its N terminal 40 amino acids[1] a sequence window yielding a high *Signify* score (>5) then this is likely to be a cleaved N terminal signal sequence *unless* the same region is also identified as part of a potential transmembrane domain. In this latter (and frequent) case, satisfactory discrimination between cleaved and uncleaved signal sequences is not possible using existing predictive methods, though it is worth bearing in

1 Eukaryotic signal peptides are seldom longer than 30 residues, and never exceed 40 (vonHeijne, 1985).

mind that cleaved N terminal signal peptides are much more common than N terminal signal anchors. If transmembrane domains are identified in the absence of a cleavable N terminal signal sequence then the most N terminal of these is predicted to be a signal anchor. Differentiation between type II ($N_{cytoplasmic}$, $C_{extracytoplasmic}$) and type III ($N_{extracytoplasmic}$, $C_{cytoplasmic}$) signal anchors is discussed below (§5.3).

A flow diagram summarising how the secreted or transmembrane nature of a novel protein sequence of interest might be examined is presented in Figure 5.3.

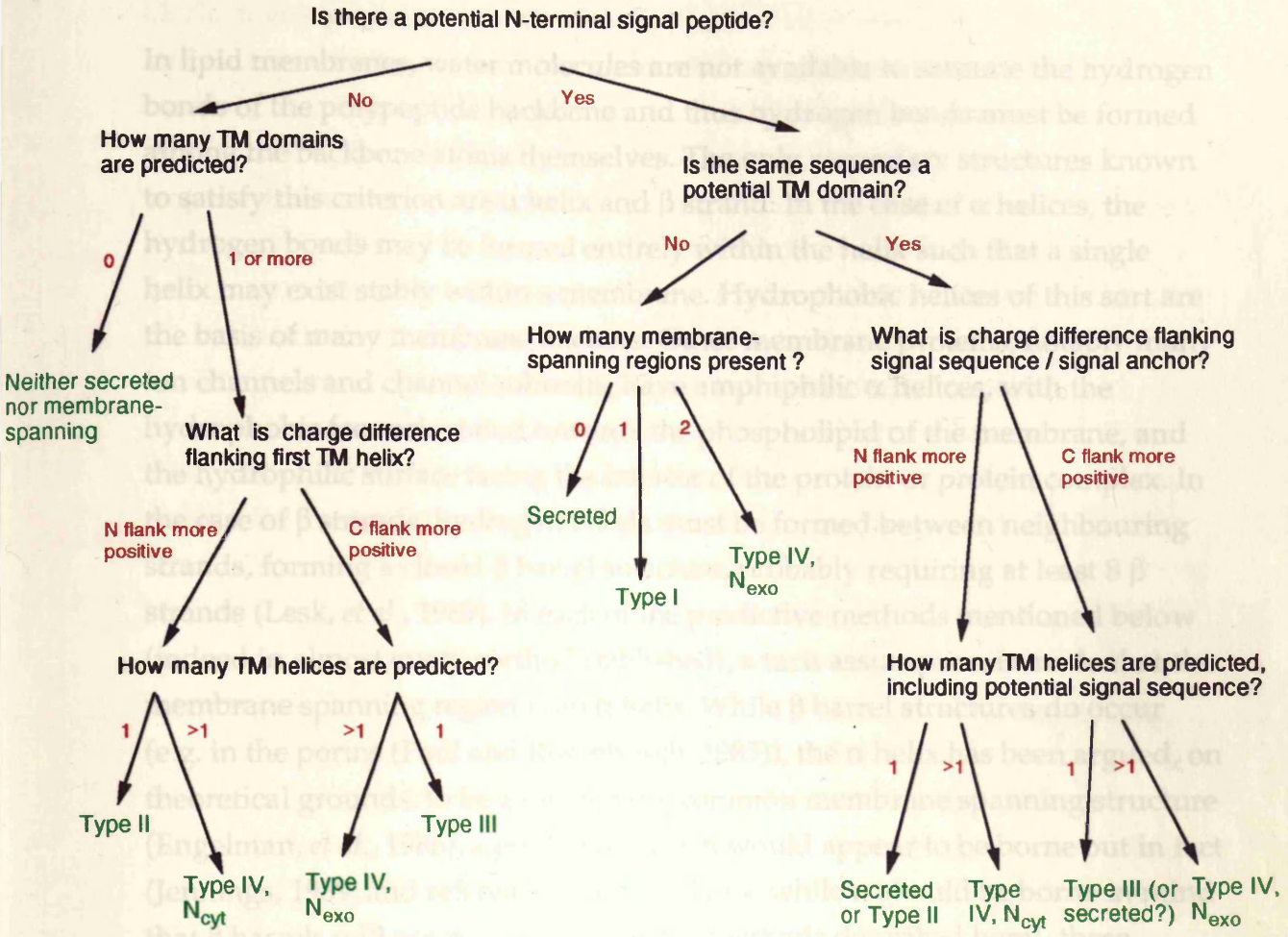


Figure 5.3: Assessment of possible transmembrane topologies of protein sequences. Choices are shown in red text; outcomes are shown in green. See text for details.

5.2 IDENTIFICATION OF MEMBRANE SPANNING DOMAINS

In lipid membranes, water molecules are not available to saturate the hydrogen bonds of the polypeptide backbone and thus hydrogen bonds must be formed among the backbone atoms themselves. The only secondary structures known to satisfy this criterion are α helix and β strand. In the case of α helices, the hydrogen bonds may be formed entirely within the helix such that a single helix may exist stably within a membrane. Hydrophobic helices of this sort are the basis of many membrane anchors. Other membrane proteins, notably many ion channels and channel subunits, have amphiphilic α helices, with the hydrophobic face orientated towards the phospholipid of the membrane, and the hydrophilic surface facing the interior of the protein or protein complex. In the case of β strands, hydrogen bonds must be formed between neighbouring strands, forming a closed β barrel structure, probably requiring at least 8 β strands (Lesk, *et al.*, 1989). In each of the predictive methods mentioned below (indeed in almost every method published), a tacit assumption is made that the membrane spanning region is an α helix. While β barrel structures do occur (e.g. in the porins (Paul and Rosenbusch, 1985)), the α helix has been argued, on theoretical grounds, to be a much more common membrane spanning structure (Engelman, *et al.*, 1986), a prediction which would appear to be borne out in fact (Jennings, 1989; and references therein). Thus, while it should be borne in mind that β barrels will not be identified by the methods described here[1], these structures are likely to be relatively rare.

Many methods for the prediction of membrane α helices within protein sequences have been proposed (Bangham, 1988; Barrantes, 1975; Biou, *et al.*, 1988; Eisenberg, *et al.*, 1984; Engelman, *et al.*, 1986; Finer-Moore and Stroud, 1984; Klein, *et al.*, 1985; Kyte and Doolittle, 1982; Rao and Argos, 1986; Rose, *et al.*, 1985; Vogel, *et al.*, 1985). Most such methods follow a common theme. The sequence in question is first scanned with a moving window to estimate the hydrophobicity of each region of the primary sequence, then those nonoverlapping windows which exceed a threshold hydrophobicity are identified as putative membrane spanning domains (Engelman, *et al.*, 1986).

---

1 Because only ten or so residues with few obvious sequence characteristics are required to span a membrane as an extended β strand, such structures are devilishly hard to identify using theoretical predictive techniques. At least one algorithm for β barrel prediction has been published (Jähnig, 1990) but its success is hard to quantify and it will therefore not be considered here.

Different prediction methods vary principally in the hydropathy values assigned to each amino acid and in the size and shape of the scanning window. Of the many available predictive methods, five have been used in this work, and each will be briefly discussed. The method in most common usage in current literature is that of Kyte and Doolittle (1982). This technique provides a benchmark against which most, if not all, potential membrane spanning proteins are tested, thus it acquires an importance somewhat out of proportion with its predictive efficacy. In conjunction with the Kyte-Doolittle (KD) technique, I will discuss the method of Engelman and coworkers (Engelman, *et al.*, 1986), which employs a different hydrophobicity scale, and one which has been cogently argued to be more biologically relevant. The same hydrophobicity scale is employed in the method of von Heijne (von Heijne, 1992), though in this latter case a trapezoidal, rather than a rectangular sliding window is used. The method of Klein and coworkers (Klein, *et al.*, 1985) was used because it performed best of the membrane spanning predictions assessed by Fasman and Gilbert (Fasman and Gilbert, 1990), while the hydrophobic moment technique of Eisenberg and colleagues (Eisenberg, *et al.*, 1984) is of interest due to its professed ability not only to identify membrane spanning regions, but also to distinguish between discrete membrane spanning domains and those likely to be associated with other helices within the bilayer.

5.2.1 The Kyte-Doolittle (KD) method

The method of Kyte and Doolittle (Kyte and Doolittle, 1982) employs a moving window which scans through a peptide sequence determining a *hydropathy*[1] value for each position of the window, evaluated as the mean hydrophobicity of the residues within the window. In analyses of this kind, the hydrophobicity attributed to each amino acid residue is obviously of paramount importance. While previous scales had been derived from partition coefficients of the various amino acids between ethanol and water (e.g. Nozaki and Tanford, 1971), Kyte and Doolittle argued that ethanol has 'unpredictable peculiarites' which make such a scale unsuitable. Instead the KD scale was derived from various combinations of water to vapour transfer free energies and the relative distributions of different residues exposed on protein surfaces or buried within their interiors (Chothia, 1976), gently massaged with a handful of subjective adjustments. Kyte and Doolittle found that the best discrimination between

1 This term has been in common usage ever since it was invented by Kyte and Doolittle. It means "strong feeling about water"

membrane spanning proteins and their globular counterparts ocurred when a window of 19 residues was used. Mean hydrophobicity in excess of 1.6 across a 19 residue window was found to correlate well with the window being contained within a membrane spanning domain. For a comparative assessment of this and the other transmembrane helix prediction algorithms used here, see §5.2.6.

5.2.2 The method of Engelman *et al*

The prediction method employed by Engelman and coworkers (Engelman, *et al.*, 1986) was in many ways similar to the KD approach, except that a different scale, that of Goldman, Engelman and Steitz (GES), was employed. While the KD method estimates only the relative hydrophobicities of *isolated* amino acid residues, the great strength of the GES scale is that it estimates the hydrophobicities of each residue when considered *as part of an α helix*. In order to derive the GES scale, the authors started with the assumption that each amino acid side chain component may be considered separately. Hydrophilic and hydrophobic groups were then assigned a free energy of transfer from water to oil. Moving a charged group from water to oil is much less energetically favourable than first neutralising the charge by protonation or deprotonation (Engelman and Steitz, 1981; Honig and Hubbell, 1984). The GES scale therefore has included in the hydrophilic component of each amino acid value, a term to reflect the energy required for protonation or deprotonation at pH 7.0. Furthermore, because oppositely charged groups arranged one turn of the helix apart may interact, and the energy required to transfer a charge pair of this kind is much less than that required to transfer two isolated charges (Honig and Hubbell, 1984; Honig, *et al.*, 1986), -10 kcal/mol is added to the estimated free energy of transfer from water to oil when paired amino and carboxyl groups are arranged at 1,4 or 1,5 positions along the helix.

Because serine and threonine side chains are known to hydrogen bond to backbone carbonyl groups (Gray and Matthews, 1984), transfer of these residues into a membrane in an α helical context is much more favourable than their polar natures would suggest. The GES scale takes this phenomenon into account in its estimates of the transfer energies of serine and threonine.

The scanning procedure of Engelman and colleagues employs a rectangular, 20 residue moving window. If the summed free energy of transfer from water to

oil of any sequence window is lower than -20kcal/mol, then the region is considered to be membrane spanning. While the GES scale is broadly similar to the KD scale, there are some significant differences. The intelligent consideration of the behaviour of amino acid residues within α helices makes the GES approach appealing, as do the specific examples provided by Engelman *et al* (1986) where the GES scale outperforms the KD scale. The performance of this technique is further assessed in §5.2.6.

5.2.3 The method of Klein *et al*

The method of Klein, Kanehisa and DeLisi (Klein, *et al.*, 1985) relies upon the Kyte-Doolittle (KD) hydrophobicity scale. This technique uses discriminant analysis to distinguish between integral membrane proteins and proteins that are extrinsic to the membrane. The approach is a beautiful example of empirical biocomputing. Imagine a function that can be applied to a peptide sequence to return a numerical value characteristic of that sequence. The function might be simple and straightforward such as the mean hydrophobicity of the protein according to a particular hydrophobicity scale, or it might be complex and subtle, obtained e.g. by Fourier analysis of arcane sequence properties. Many such functions will return values that in no way correlate with whether a protein is integral to a membrane or extrinsic to it but some (such as that used by Kyte and Doolittle, above) would be expected to show such a correlation. Let us denote P(I) and P(E) as the probabilities that any random protein is intrinsic or extrinsic to a membrane, these probabitilies being estimated from the relative abundances of each group within proteins as a whole. Let us also declare P(I/x) and P(E/x) as the conditional probabilities that a protein is integral or extrinsic to the plasma membrane given that it returns a value of x for a function applied to the sequence. A sequence is therefore assigned to the extrinsic group if:

$$P(E/x) > P(I/x) \qquad (1)$$

but, according to Bayes' theorem:

$$P(I/x) = \frac{P(x/I).P(I)}{P(x/I).P(I)+P(x/E).P(E)}$$

and

$$P(E/x) = \frac{P(x/E).P(E)}{P(x/I).P(I)+P(x/E).P(E)} \qquad (2)$$

making (1) equivalent to :

$$P(E).P(x/E) > P(I).P(x/I) \qquad (3)$$

If large groups of extrinsic and intrinsic proteins are available to test, P(x/I) and P(x/E) can be derived from estimates of the probability distributions of x in each group of proteins.

These probabilities can then be used to formulate a *discriminant function*, which, when applied to x, will distinguish between integral and extrinsic proteins. The laws of multivariate statistics decree that such a function will be either linear or quadratic, depending on the assumptions made about the variances of P(x/E) and P(x/I). In fact, Klein *et al* presented both a linear and a quadratic function, which operate on the mean hydrophobicity of the most hydrophobic 17 residue window ($x=H_{max}[17]$), according to the KD scale. These discriminant functions (specific examples of inequality (3)) were as follows:

P(E/x) > P(I/x) if:

Linear function: $-9.02x + 14.27 > 0$ (4)[1]

Quadratic function: $1.05\ x^2 - 12.30x + 17.49 > 0$ (5)

When these functions are applied back to the set of proteins from which the probability distributions were estimated, the quadratic function misclassifies none of the proteins of the 102 member training set, while the linear function misclassifies only one. When applied to a test set distinct from the training set,

1 Note that (4) is equivalent to $x<1.58$, very close indeed to the independently derived value of 1.6 used by Kyte and Doolittle.

the linear discriminant function misclassified between 1% and 2% of proteins, though it should be noted that some entries in the test set were homologous to some of the training set proteins, thus, in the mathematical poetry of Klein and colleagues, the test set is 'not fully orthogonal to the training set'.

A great advantage stems from the statistical rigour of the KKD procedure. Whether the linear or quadratic discriminant function is used, the odds (P(E/x):P(I/x)) are expressed by e^b, where b is the left hand side of the inequality in (4) or (5).

Klein and coworkers also addressed the question of the exact position of the boundaries of membrane spanning domains. They defined inner and outer boundary positions on either side of the helix and argued that the actual boundary was likely to be the average between the two. Inner boundaries were set at each end of the 17 residue window which showed a hydrophobicity maximum, and outer boundaries were taken as the outermost residue of the last window to be classified as integral by the discriminant function, as the analysis window moves away from the hydrophobic core. Although only a small number of membrane helix boundaries had been determined experimentally in 1985, the success of this boundary prediction led the authors to declare that '... the residues at the boundary of a membrane-spanning segment are predictable to within the error inherent in the concept of boundary.'

A Macintosh implementation of the membrane spanning helix prediction method of Klein *et al* has been prepared by A. Luettke and P. Markiewicz, and is used in this work. Their program is called *Plota_KKD*, and is available from the EMBL file server (Fuchs, *et al.*, 1990). While this program is helpful in calculating the Kyte-Doolittle hydrophobicities, it generates erroneous results for the linear and quadratic discriminant functions, and should be used with care.

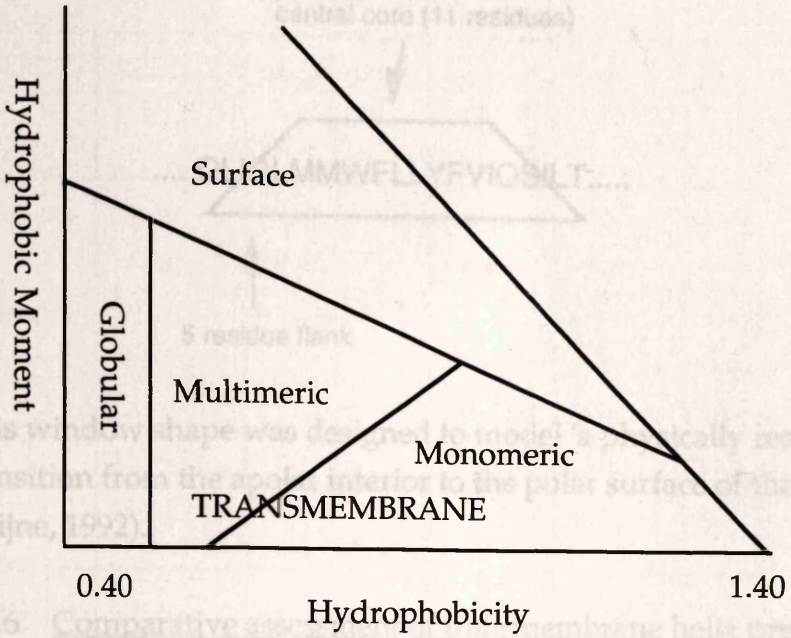
5.2.4 The method of Eisenberg *et al*

The membrane helix prediction method of Eisenberg and coworkers (Eisenberg, *et al.*, 1984) was designed both to identify membrane spanning helices within amino acid sequences and also to determine whether they are amphiphilic. A lone transmembrane domain surrounded on all sides by

phospholipid molecules would be expected to be hydrophobic on every face, and thus to have little amphiphilicity. Conversely, a protein lying on the exterior of a membrane, or one interacting within the membrane with other membrane spanning helices, might be highly amphiphilic. Thus determination of the amphiphilicity of a potential membrane spanning sequence could reveal some interesting hints as to its function.

The Eisenberg technique has features common to the KD and GES methods. Again, a moving window (this time of 21 residues) is run across the amino acid sequence, and a hydrophobicity is assigned to each sequence segment. In this instance, the hydrophobicity scale used is a consensus scale based on five other scales (Eisenberg, *et al.*, 1982b). Windows are initially assigned as candidate membrane spanning regions if their mean hydrophobicity exceeds 0.42 (a value between cysteine and glycine). However, all membrane spanning candidate regions are rejected unless either one segment has a mean hydrophobicity greater than 0.68 (between alanine and methionine) or two segments exist whose summed mean hydrophibicities exceed 1.10. This stipulation reflects the proposal that highly hydrophobic *initiator helices* or helix pairs are required to allow the initial folding ot the protein into the membrane. If either criterion is fulfilled then all nonoverlapping segments with mean hydrophobicities greater than 0.42 are initially accepted as putative membrane spanning domains.

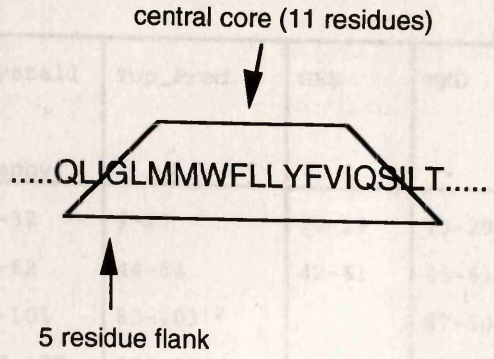
Once potential membrane spanning helices have been determined in this way, the amphiphilic character of each is assessed by determining its helical hydrophobic moment (Eisenberg, *et al.*, 1982a). The helical hydrophobic moment is simply a measure of the amphiphilicity of an α helix. The angle between one side chain and the next is taken as 100°, the value for an idealised α helix. Eisenberg and coworkers observed that plotting the hydrophobic moment of the 11 residue subwindow with the highest hydrophobic moment against the hydrophobicity of that subwindow for each potential membrane spanning segment, allowed differentiation between helices of globular proteins and monomeric, multimeric, and membrane surface domains of membrane associated proteins:



Plota_TMH, a Macintosh implementation of the Eisenberg transmembrane helix prediction is available as part of the *Plota* package of sequence analysis programs written for Macintosh by A. Luettke and P. Markiewicz. This program is available via the EMBL file server (Fuchs, *et al.*, 1990), and has been used in this work.

5.2.5 The method of von Heijne

von Heijne's transmembrane domain prediction technique (von Heijne, 1992), like that of Engleman and colleagues (§5.2.2) uses the GES hydrophobicity scale (Engelman, *et al.*, 1986). However, while the Engleman method assesses the mean hydrophobicity of windows of twenty amino acids, the von Heijne technique calculates a weighted mean, with the residues towards the centre of the window contributing more to the hydrophobicity score than the peripheral residues. This weighting method is referred to as a trapezoidal sliding window. The total length of the window used was 21 amino acids, with 11 residues in the central core, and five on each tapering flank:



This window shape was designed to model 'a physically reasonable, soft transition from the apolar interior to the polar surface of the membrane' (von Heijne, 1992).

5.2.6 Comparative assessment of transmembrane helix predictions

Assessments of the relative efficacies of transmembrane domain prediction techniques have been made by applying the predictive algorithms to proteins whose crystal structures are known (Fasman and Gilbert, 1990; Jähnig, 1990). In a similar vein, I here include a brief comparative assessment of the five methods used in this work. These observations extend the analysis of Fasman and Gilbert, firstly by including two algorithms not considered in their paper and secondly by adding two more membrane proteins. The structure of one of these (bacteriorhodopsin) has not been solved to quite the same resolution as the others, but is sufficiently well characterised to accurately place its seven membrane spanning α helices (Henderson, *et al.*, 1990). The other new membrane protein is the sheep prostaglandin G/H synthase enzyme. One original report of the primary sequence of this protein proposed a central transmembrane helix (Merlie, *et al.*, 1988), while a different report of the same sequence did not (DeWitt and Smith, 1988), demonstrating that this is a difficult to classify, borderline case. The crystal structure of the enzyme was reported recently (Picot, *et al.*, 1994), showing that, while it is strongly membrane bound, the protein has no membrane spanning domains (Picot and Garavito, 1994). The purpose of its inclusion was therefore to judge the extent to which the different algorithms might be tempted into declaring false positives. The other proteins considered are the three reaction centre proteins of *Rhodopseudomonas viridis* (Deisenhofer, *et al.*, 1985), as used by Fasman and Gilbert.

Table 5.1, below, shows the performance of the five prediction algorithms when applied to the membrane proteins listed above.

	Crystallography	*Top_Pred*	GES	KKD	Eisenberg	Kyte-Doolittle
Bacterio-rhodopsin	10-32 38-62 80-101 108-127 136-157 167-193 203-227	9-29 44-64 83-103 [2] 106-126 137-157 175-195 203-223 [3]	10-29 42-61 108-127 136-155 178-197 204-223	13-29 45-61 87-103 108-124 139-155 176-192[1] 206-222	9-29 83-103 107-127 135-155 178-198 202-222	11-29 44-62 108-126 135-153 205-223
H chain	12-37	12-32	12-31	16-32	12-32	12-30
M chain	52-78 110-139 142-167 197-225 259-285	55-75 110-130 145-165 200-220 267-287	53-72 111-130 149-168 207-226 270-289	56-72 151-167 210-226 270-286	52-72 110-130 139-159 205-225 270-290	54-72 149-167 207-225 272-290
L chain	32-55 84-112 115-140 170-199 225-251	30-50 84-104 115-135 138-158 [4] 175-195 233-253	31-50 84-103 112-132 179-198 233-252	34-50 118-134 180-196 235-251[1]	31-51 112-132 178-198 233-253	32-50 122-140 180-198 233-251
PG synthase		8-28 [7] 77-97 [4] 191-211 [5] 287-307 [6]	6-25 285-304			

Table 5.1: Performance of transmembrane domain prediction algorithms when applied to membrane proteins of known structure.

[1] Only the linear discriminant function acknowledges this helix. The quadratic function fails to do so.
[2] A possible score of 0.90 is generated here.
[3] A possible score of 0.83 is generated here in the absence of ion pair stabilisation energy. With 5.0 kcal/mol charge pair stabilisation, the score rises to the 'definite' 1.15
[4] A possible score of 0.71 is generated here.

[5] A possible score of 0.66 is generated here.
[6] A 'definite' score of 1.21 is generated here.
[7] A 'definite' score of 1.16 is generated in this region, which corresponds to the cleaved N terminal signal peptide (see §5.1.3).

Three scores are derived for the performance of each method, and are shown in table 5.2. Firstly, any predicted helix which overlaps by 10 residues or more with a proven transmembrane helix is said to identify that helix. The sensitivity score is the number of helices correctly identified as a percentage of the total number present. The selectivity score is the number of overpredictions made, thus a score of zero is optimal. Note that overpredictions of the cleaved signal peptide of P/G synthase are not included in this table. This form of misclassification is a separate case, which has been considered in §5.1.3. The alignment score is the average distance of the central residue of each predicted helix from the centre of the actual helix.

Prediction method	Sensitivity (%) (overpredictions)	Selectivity (overpred'ns)	Alignment
Top_Pred cutoff 0.6; 0kcal/mol ion pair stabilisation (IPS) energy	100	4	2.58
Top_Pred cutoff 0.6; 5.0kcal/mol IPS energy	100	4	2.58
Top_Pred cutoff 1.0; 0kcal/mol IPS energy	89	1	2.62
Top_Pred cutoff 1.0; 5.0kcal/mol IPS energy	94	1	2.59
GES 0kcal/mol IPS energy	94	1	3.74
GES 10 kcal/mol IPS energy	94	1	3.74
KKD linear function	89	0	2.81
KKD quadratic function	77	0	2.73

Kyte-Doolittle	77	0	3.18
Eisenberg	89	0	3.84

Table 5.2 Assessment of sensitivity, selectivity, and alignment accuracy of transmembrane domain predictions. Data are from table 5.1.

Despite the small size of the data set used, an important pattern begins to emerge. The different techniques vary both in their sensitivities and selectivities. The highly selective techniques, with no overpredictions, e.g. the KD and Eisenberg methods, lose out in sensitivity, failing to predict 4 and 2 transmembrane domains respectively. At the other end of the spectrum, the highly sensitive *Top_Pred* technique, using a cutoff score of 0.6, alone identifies every *bona fide* transmembrane domain, but is insufficiently selective, giving 4 overpredictions. As a first approximation, the methods based on the GES scale (Engleman, *Top_Pred*) are overpredictors, while those based on the Kyte-Doolitle (KD, KKD) and Eisenberg's consensus scale (Eisenberg) are underpredictors. When the GES scale is used, inclusion of a favourable ion pair stabilisation energy (5 to 10 kcal/mol) appears to enhance sensitivity without a detrimental effect on selectivity. The *Top_Pred* technique displays the most accurate positioning of helices, while the Eisenberg technique is rather wayward in this respect.

5.3 PREDICTING TRANSMEMBRANE PROTEIN TOPOLOGY: *Top_Pred*

Before going on to describe how the putative protein products of the *shaking-B* locus and their homologues perform when subjected to the scrutiny of the signal peptide and transmembrane helix prediction algorithms described above, one last class of prediction algorithms is worth mentioning, this being the group of techniques devised to predict the *topology* of transmembrane helices: i.e. which end of each membrane spanning helix faces in and which end faces out of the cytoplasm. Two related techniques will be discussed (Hartmann, *et al.*, 1989; Sipos and von Heijne, 1993).

The method of Hartmann and coworkers (Hartmann, *et al.*, 1989), focuses on the problem of determining the orientation of proteins containing uncleaved signal anchor sequences. These authors noted a strong correlation between the

charge difference between the 15 residues immediately N terminal and the 15 residues C terminal to the signal anchor sequence, the region on the cytoplasmic side of the membrane containing more positive charges than that on the extracytoplasmic side. These authors proposed that simple observation of charge differences between the regions N terminal and C terminal to the signal anchor thus enables accurate prediction of the topology of the protein, with each successive membrane spanning segment in type IV proteins having the opposite orientation to the previous transmembrane domain. In the set of 91 proteins considered, the charge difference method incorrectly predicted the orientation of only 3 proteins.

An extension of this approach was used by Sipos and von Heijne (Sipos and von Heijne, 1993) to predict the topology of type IV eukaryotic membrane proteins. In addition to confirming the validity of the *charge difference rule* of Hartmann and colleagues, these authors demonstrated that for short, polar segments intervening between membrane spanning stretches in multipass proteins, the distribution of arginines and lysines is biased towards the cytoplasmic loops- the so called *positive inside* rule (see also Nakashima and Nishikawa, 1992). If the segments intervening between membrane spanning regions are longer than 60 residues, the positive inside rule breaks down, thus, while it is a useful tool in predicting the topologies of some proteins, it cannot be applied to those in which most nonmembrane regions are long.

Top_Pred, a Macintosh program incorporating the transmembrane helix prediction method of von Heijne (von Heijne, 1992) and the topology prediction algorithms of Sipos and von Heijne (Sipos and von Heijne, 1993) has been prepared by these authors, and is used in this work. This program is available from the EMBL file server (Fuchs, *et al.*, 1990). Users should, however, be warned, that version 3.0 of this program somewhat surreally ignores all tryptophan residues in the input sequence and also generates incorrect transmembrane helix prediciton scores. This shortcoming has been remedied in the recent version 3.2 of the program which was a generous gift from Gunnar von Heijne.

5.4 RESULTS OF SIGNAL PEPTIDE AND MEMBRANE TOPOLOGY PREDICTIONS

In the light of the above considerations of signal peptide and transmembrane helix predictions, I will attempt to model the membrane topologies of the Shaking-B proteins proposed in chapters 3 and 4. I will initially consider only the Shak-B(44.4) protein; the other proposed species are largely contained within this sequence, and will be discussed after a topology model for their larger relative has been presented.

I will also consider the application of signal peptide and transmembrane topology predictions to the neural ShakB(42.9) sequence, and to the sequences of the Shak-B homologues Ogre and Unc-7. On the basis of their primary sequences, it has been proposed by others that Shak-B(42.9) and Ogre each contain a single membrane spanning α helix, while Unc-7 has been postulated to have as many as four transmembrane helices (Krishnan, *et al.*, 1993; Starich, *et al.*, 1993; Watanabe and Kankel, 1990). Each of the transmembrane domains proposed within Unc-7 lies in the region of homology with Shak-B and Ogre proteins. In light of the highly significant primary sequence homology among these proteins (Figure 4.9) such radically different transmembrane topologies seems questionable. A critical reassessment of the possible transmembrane dispositions of the Shak-B(42.9), Ogre and Unc-7 proteins is therefore also included here.

5.4.1 Prediction of signal peptides

As discussed above, the most accurate method for prediction of cleaved N terminal signal peptides is that of von Heijne (von Heijne, 1986), as implemented in the *Signify* program. Recall that any window contained within the N terminal 40 residues of a protein whose score exceeds 5 is considered to be a likely signal peptide, while a window whose score exceeds 2 is regarded as possible (§5.1.1). When applied to the Shak-B(44.4), Ogre and Unc-7 sequences, no N terminal sequences resembling cleaved signal peptides are detected. Both ShakB(44.4) and Ogre do contain N terminal regions which give high *Signify* scores: Shak-B(44.4) scores 3.65 for the potential cleavage between residues 43 and 44 while Ogre scores 5.20, also for cleavage between amino acids 43 and 44. However in both cases these regions are substantially further from the N terminus than would be expected for a eukaryotic signal peptide [1].

[1] In Shak-B(44.4), a score of 3.64 is generated for a sequence whose cleavage site would be between residues 42 and 43, the largest known eukaryotic signal peptide being 35 amino acids in length (von Heijne, 1986). In Ogre, a signal peptide of the same length would generate a score of 5.19.

Furthermore, at an almost exactly homologous position in the Unc-7 protein, a score of 5.98 is generated, even though this cleavage position would be between residues 163 and 164, and thus does not represent a signal peptide cleavage (see Figure 5.4). As mentioned above (§5.1.3), signal anchor sequences may also generate high *Signify* scores, and indeed transmembrane helix predictions (see below) identify these regions generating high *Signify* scores as strong candidate transmembrane domains.

The putative signal peptide in Shak-B(42.9) is harder to assess. This was proposed as a genuine signal peptide by Krishnan and coworkers (Krishnan, *et al.*, 1993). Figure 5.5 shows the *Signify* screen output for the Shak-B(42.9) protein. The window beginning at position 20 generates a score greater than 8, thus identifying a strong signal peptide candidate. The predicted cleavage site in Shak-B(42.9) both generates a higher *Signify* score than those of its homologues and anticipates a 32 residue signal peptide which, while longer than average, is within the range for which precedents exist (von Heijne, 1985). However, the transmembrane helix prediction results described below anticipate that the putative signal peptide region of Shak-B(42.9), if uncleaved, would contribute to a transmembrane helix, covering residues 14-34, thus this region may be a signal anchor. The highest scoring cleavage site coincides almost exactly with those present in the homologous sequences (Figure 5.4), which are probably spurious. On the other hand in a number of instances, type II signal anchor domains have been converted into cleaved signal peptides by creating deletions N terminal to the cleavage site (Lipp and Dobberstein, 1986; Schmid and Spiess, 1988), though N terminal deletions of type II proteins do not always elicit signal peptide cleavage (e.g. Hegner, *et al.*, 1992; Hong and Doyle, 1990). There exists, therefore, the interesting possibility that following duplication of ancestral signal anchor containing exons, the N terminus of the Shak-B(42.9) protein has evolved a cleaved N terminal signal sequence.

The antimorphic *shak-B*^{Passover} allele is caused by a C to T transition generating an arginine to tryptophan substitution at residue 33 of the Shak-B(42.9) protein (Krishnan, *et al.*, 1993). This substitution has been proposed to impair signal peptide cleavage (the putative cleavage site being between residues 32 and 33) so leaving a persistent preprotein whose nonfunctional interactions with other molecules underpin the antimorphic behaviour of the allele. This is an elegant proposal, but is not the only possible explanation.

Could this substitution reasonably be anticipated to impair signal peptide cleavage? Nothwehr and colleagues (Nothwehr, *et al.*, 1990) analysed the effects of amino acid substitutions at the +1 position relative to the signal peptidase cleavage site of a human preprotein pre(Δpro)apo-A-II. The normal residue in this position is lysine (compare arginine in the potential +1 position of Shak-B(42.9)). Substitution with seven other residues was found not to effect the *level* of signal peptidase cleavage observed, but have a slight influence on the cleavage *site* chosen. The substitutions were found to increase, to different degrees, the use of an alternative cleavage site 2 residues N terminal to the normal site. In pre(Δpro)apo-A-II, however, the alternative -2 site is predicted to be highly favourable, having a *Signify* score of 6.26, compared to 6.65 for the normal, wild type site. In Shak-B(42.9) there is no strong candidate for an alternative cleavage site in the *shak-B*^{Passover} mutant. The proposed optimal cleavage site is a predicted to be less favourable in the *shak-B*^{Passover} mutant (scoring 7.74) than in the wild type sequence (scoring 8.09), so it seems at least possible that the *shak-B*^{Passover} mutation, in the absence of an alternative cleavage site, prevents cleavage of some of the Shak-B(42.9) protein. However, given that the mutant cleavage site still appears highly favourable, it is also possible that the preprotein is still cleaved normally, and that it is the alteration of the first residue of the mature protein that is responsible for the mutant phenotype.

Under the alternative signal anchor model, the R to W substitution lies at the extracytosolic end of the first transmembrane helix of Shak-B(42.9) (see below). It is thus possible that the *shak-B*^{Passover} lesion might instead exert its effect by altering the properties of this transmembrane domain. Distinguishing between the cleaved signal peptide and uncleaved signal anchor alternatives for Shak-B(42.9) is not currently possible and must await biochemical investigations (see chapter 6).

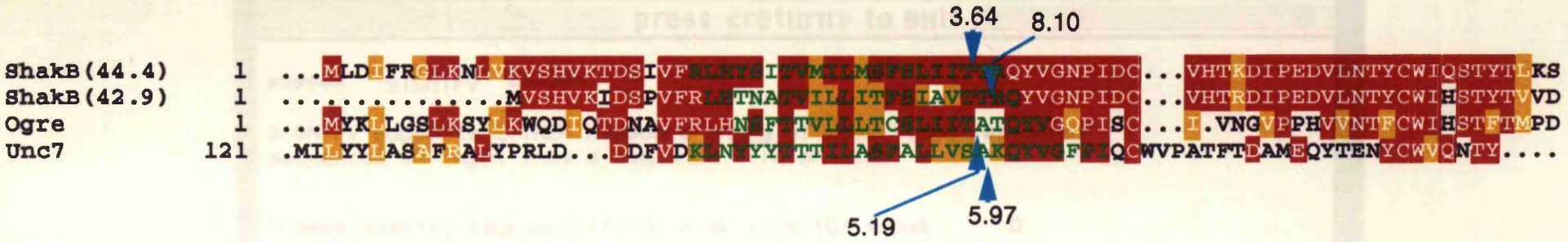


Figure 5.4: *Signify* maxima in the N terminus of Shak-B(44.4) and homologous regions of Shak-B(42.9), Ogre and Unc-7. A PILEUP multiple sequence alignment is shown. Identical residues are boxed in red, chemically conservative replacements are shown in amber. Blue arrows indicate the positions generating high *Signify* scores; the exact scores for each sequence are shown. Residues in green are those predicted by transmembrane prediction algorithms to potentially form transmembrane signal anchors (see §5.1.4.b).

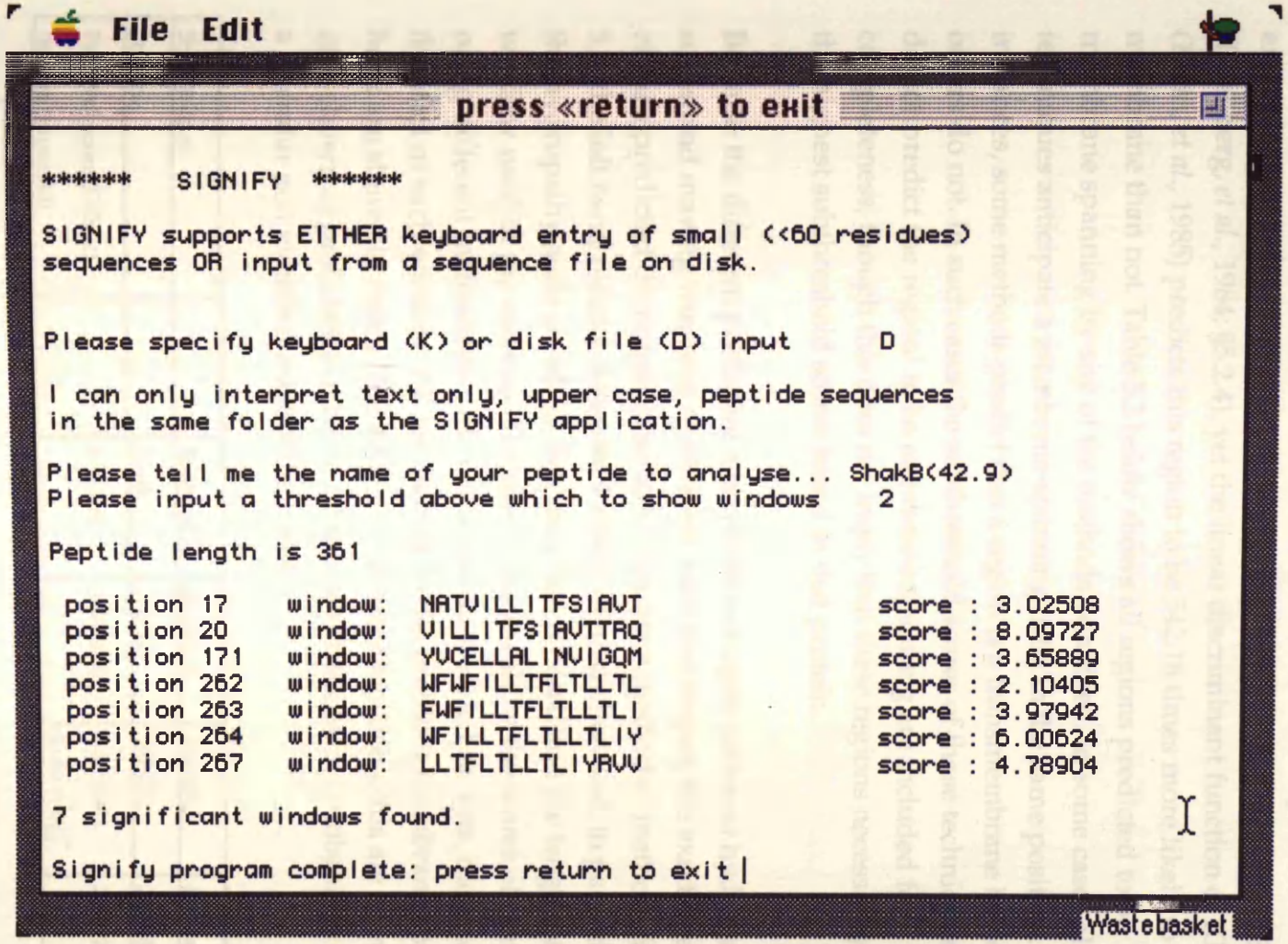


Figure 5.5: Screen output of the *Signify* program searching for signal peptides in Shak-B(42.9). The high scoring window starting at position 20 highlights the potential N terminal signal peptide cleavage site between residues 32 and 33.

5.4.2 Prediction of membrane spanning domains

As discussed above, five membrane helix predictions, based upon three different hydropathy scales, have been used to model the number and positions of transmembrane helices in each protein. These different methods yield strikingly different results. The 405-421 region of the Unc-7 protein, for example, is not identified as a transmembrane helix by the method of Eisenberg (Eisenberg, *et al.*, 1984; §5.2.4), yet the linear discriminant function of Klein *et al* (Klein, *et al.*, 1985) predicts this region to be 542.18 times more likely to span a membrane than not. Table 5.3 below shows all regions predicted to be membrane spanning by any of the methods described. In some cases all techniques anticipate a membrane-spanning helix in the same position. In other instances, some methods predict that a region is a transmembrane helix while others do not. In such cases the subthreshold scores of those techniques which do not predict the region[1] to be membrane spanning are included for completeness, though this does not imply that these regions necessarily yield the highest subthreshold scores found in that protein.

Because the different prediction algorithms rely upon different hydropathy scales and moving windows of different sizes and shapes, the exact boundaries of each predicted transmembrane helix vary from method to method. In table 5.3, the left hand column states the method of prediction and, in parentheses, the hydropathy scale on which the prediction is based and the length of window used in the analysis. The other columns show the scores obtained using different methods on different windows, and, in brackets, the position of the start of each window. All results that do not predict a membrane spanning helix are shown in italics. The KKD discriminant function odds are in every case shown as probability of being a transmembrane helix : probability of being a globular non membrane-spanning region.

Shak-B(44.4)		Region A	Region B	Region C	Region D
Position		22-45	108-129	181-203	268-288
Top_Pred score (GES, 21 residue trapezoid)		1.79 (26)	1.40 (109)	0.80 (181) [0.89 (183)]*	2.61 (268)

1 In these cases, the subthreshold window chosen is the highest scoring window that overlaps with the window(s) specified as membrane spanning helices by other algorithms.

GES free energy-kcal/mol (GES, 20)		-27.2 (22)	*-15.5 (108)*	*-14.4 (181)*	-53.5 (268)
Mean Hydrophobicity (KD, 19)		1.64 (26)	*1.06 (109)*	*1.58 (181)*	2.49 (270)
KKD linear odds (KD, 17)		19.26:1 (26)	*1:8.11 (110)*	19.26:1 (183)	1019.43:1 (272)
KKD quadratic odds (KD, 17)		8.78:1 (26)	*1:14.99 (110)*	8.78:1 (183)	274.89:1 (272)
Eisenberg algorithm (Consensus scale, 21; hydrophobic moment, 11)		*Globular (25)*	*Globular (108)*	*Surface seeking (183)*	Multimeric transmembrane (268)
Shak-B(42.9)		Region A	Region B	Region C	Region D
Position		11-34	97-120	170-192	257-277
Top_Pred score (GES, 21 residue trapezoid)		1.42 (14)	1.32 (98)	0.80 (170) [0.89 (172)]	2.61 (257)
GES free energy-kcal/mol (GES, 20)		-22.8 (11)	-22.9(101)	*-14.4 (170)*	-53.5 (257)
Mean Hydrophobicity (KD, 19)		*1.51 (14)*	*0.86 (98)*	*1.58 (170)*	2.49 (259)
KKD linear odds (KD, 17)		5.45:1 (14)	*1:64.56 (99)*	19.26:1 (172)	1019.43:1 (261)
KKD quadratic odds (KD, 17)		2.70:1 (14)	*1:153.1 (99)*	8.78:1 (172)	275:1 (261)
Eisenberg algorithm (Consensus scale, 21; hydrophobic moment, 11)		*Globular (14)*	*Globular (97)*	*Surface seeking (172)*	Multimeric transmembrane (257)
Ogre	Region A	Region F	Region B	Region C	Region D
Position	28-48	58-78	106-126	178-200	268-289
Top_Pred score (GES, 21 residue trapezoid)	1.77 (28)	0.84 (58)	1.41(106)	0.75 (180) [0.81 (180)]	2.70 (268)
GES free energy-kcal/mol (GES, 20)	-31.7 (29)	*-19.4 (58)*	-23.5 (107)	*-14.2 (180)*	-55.7 (269)
Mean Hydrophobicity (KD, 19)	1.61 (30)	*0.83 (51)*	*0.83 (106)*	*1.43 (180)*	2.89 (270)

KKD linear odds (KD, 17)	5.96:1 (29)	*1:95.6 (52)*	*1:8.87 (107)*	3.47:1 (182)	190727.31:1 (271)
KKD quadratic odds (KD, 17)	2.94:1 (29)	*1:236.5 (52)*	*1:18.05 (107)*	1.75:1 (182)	13833.59:1 (271)
Eisenberg algorithm (Consensus scale, 21; hydrophobic moment, 11)	*Globular (28)*	Multimeric transmembrane (51)	*Globular (106)*	*Surface seeking (178)*	Monomeric transmembrane (268)
Unc-7	Region E	Region A	Region B	Region C	Region D
Position	115-135	144-171	221-241	313-340	404-424
Top_Pred score (GES, 21 residue trapezoid)	0.62 (115)	1.24 (145)	1.46 (221)	0.87 (313)	2.21 (404)
GES free energy-kcal/mol (GES, 20)	*-7.3 (116)*	-21.6 (152)	-32.8 (222)	-20.3 (315) -20.3 (320)	-42.7 (405)
Mean Hydrophobicity (KD, 19)	*1.10 (117)*	*1.45(153)*	2.1 (223)	*1.53 (321)*	1.97 (404)
KKD linear odds (KD, 17)	*1:26.75 (117)*	*1:3.38(153)*	240.76:1 (225)	*1:1.22 (323)*	542.18:1 (405)
KKD quadratic odds (KD, 17)	*1:58.54 (117)*	*1:6.62(153)*	82.36:1 (225)	*1:2.36 (323)*	163.31:1 (405)
Eisenberg algorithm (Consensus scale, 21; hydrophobic moment, 11)	*Globular (115)*	Multimeric transmembrane (151)	Multimeric transmembrane (221)	*Globular (320)*	*Globular (405)*

* Values in square brackets are results when ion pair stabilisation at 5.0 kcal/mol is included

Table 5.3: Results of membrane prediction algorithms when applied to the sequences of Shak-B proteins and their homologues.

At the outset, let us consider as 'definite' transmembrane helices all those that are predicted by every algorithm to be transmembrane, i.e. region D of Shak-B(44.4), region D of Shak-B(42.9), region D of Ogre, and region B of Unc-7.

Region D of Unc-7 may safely be added to the definite list. It is emphatically predicted to be a membrane spanning helix by all methods except that of Eisenberg *et al.* Due to dependence on hydrophobic moment, the Eisenberg

algorithm rejects a 21 residue window starting at position 405, yet if this window were *less* hydrophobic, adjacent regions (e.g. the window starting at 403) would be accepted as multimeric transmembrane helices. This is a good example of the shortcoming of the Eisenberg technique. Regardless of its efficacy in predicting the character of those transmembrane helices which it does acknowledge, it is, as shown in table 5.1, an *underpredictor* of transmembrane helices. This being so, we may reasonably accept those potential transmembrane helices identified by all algorithms except that of Eisenberg *et al*, these being region A of Shak-B(44.4) and region A of Ogre.

We have thus accounted for 7 of the 18 potential TM helices. With the proposed status of most candidate helices left, for the time being, unresolved, it is informative to consider the different potential transmembrane regions in these homologous proteins, and how they relate to one another. Figure 5.6 shows once again the multiple sequence alignment among Shak-B, Ogre, and Unc-7 proteins. The candidate transmembrane regions, with the exception of region E of Unc-7and region F of Ogre, are shown in green text, and two facts are immediately obvious. The first is that these regions superimpose almost perfectly. The second is that they lie in areas of high homology. Of the 59 positions in which all four proteins are seen to have identical residues, 30 lie within these candidate transmembrane segments. Given the relative sizes of the candidate transmembrane *versus* nonmembrane regions, this implies that these potential membrane spanning domains have more than twice the concentration of four way identities compared to the remainder of the proteins[1].

---

1 While it is true that considering candidate transmembrane domains with few hydrophilic residues automatically selects regions with a reduced amino acid diversity this diversity is not reduced to such a degree that chance identities would be at all likely to create any semblance of homology in regions not evolutionarily conserved.

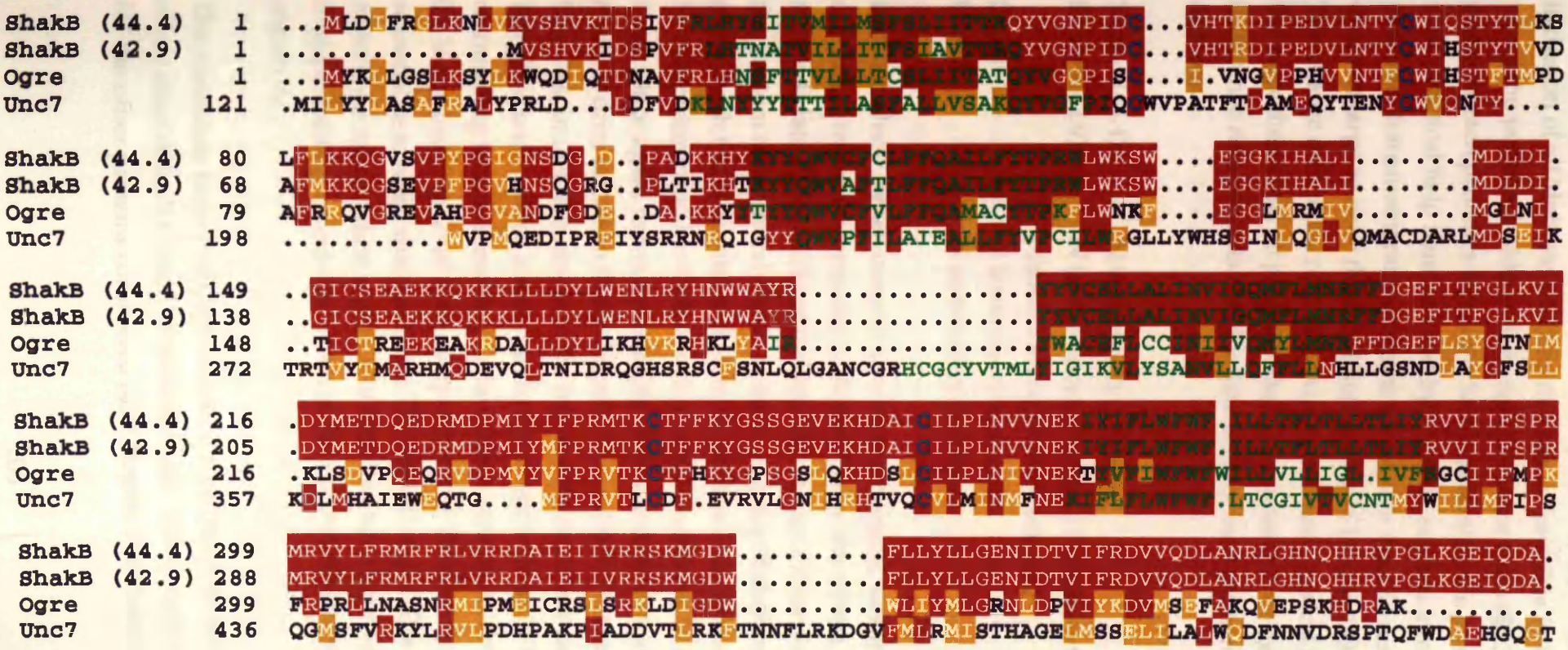


Figure 5.6: PILEUP sequence alignment among Shak-B(44.4), Shak-B(42.9), Ogre and Unc-7. Identities are highlighted in red, chemically conservative changes (S/T, M/V/I/L, W/Y/F, A/G, R/K, D/E, Q/N) are shown in orange. Conserved cysteine residues are shown in cyan text. Independently determined candidate transmembrane domains A to D are shown in green text, and are seen to coincide closely.

An ideal next step would be to test the different topologies resulting from the inclusion of all 'definite' and any number of 'possible' helices for each protein, using the positive inside and charge difference rules (§5.3) to arrive at a best guess of each protein's topology. Sadly, however, the positive inside rule is of little practical help here, as most hydrophilic regions intervening between potential transmembrane domains are greater than 60 residues in length, and thus too large to obey the rule (Sipos and von Heijne, 1993; von Heijne, 1992). The charge difference rule, however, is of more practical help. If we consider the hypothesis that all proposed helices in table 5.3 are genuine, the charge difference rule predicts that the first helix in the 4 proteins will be disposed as follows:

Shak-B(44.4):	N terminus cytoplasmic
Shak-B(42.9):	N terminus cytoplasmic (Or extracytoplasmic after signal peptide cleavage, if this occurs)
Ogre:	N terminus cytoplasmic
Unc-7:	N terminus cytoplasmic

Because Unc-7, when compared to its homologues, has an extra region E candidate transmembrane helix, N terminal to the others, the assumptions that all candidate TM helices are real and the charge difference rule holds require the protein to be entirely 'out of phase' with the Shak-B proteins and out of phase with the N terminus of Ogre. In other words, homologous regions would be cytoplasmic in Unc-7 while being extracytoplasmic in its homologues, or *vice versa*. Once again this is a possibility that cannot be completely ruled out, but *a priori*, it does not seem likely. Inspection of table 5.3 reveals that region E of Unc-7 is, in fact, the lamest of the transmembrane helix candidates, getting a *Top_Pred* score which is only just in the possible range, while being a resounding failure according to the other assessments. When a cutoff score of 0.6 is applied, *Top_Pred* is an overpredictor of TM domains, thus it seems reasonable to reject this candidate TM helix. If region E is rejected, the charge difference rule predicts region A of Unc-7 to have a cytoplasmic N-terminus, thus permitting a model in which all four proteins have similar topologies in region A.

The candidate helix F of Ogre would, if accepted, also render the predicted orientation of its B,C and D transmembrane helices opposite to those predicted for the other proteins considered here. Region F is identified as a possible

multimeric transmembrane helix by the method of Eisenberg and generates a 'possible' *Top_Pred* score of 0.84, but is rejected by the other prediction algorithms. Considered in isolation, this candidate helix, being accepted by the relatively stringent Eisenberg technique might be a fairly good candidate. However, the high level of homology of Ogre and Shak-B proteins makes it seem unlikely that a TM helix will be present at 58-78 in Ogre, but absent from the homologous regions of Shak-B(44.4) and ShakB(42.9), yet in these latter regions, there is no suggestion of a transmembrane domain. Thus, if Ogre and Shak-B(44.4) do, as their high homology suggests, have identical topologies, then the candidacy of region F of Ogre seems rather weak, and it is probably fair to reject this potential transmembrane helix.

Region A of Shak-B(42.9) is, as discussed above (§5.4.1), a good candidate N terminal signal peptide. According to the *Top_Pred*, Engelman, and KKD algorithms, it is also a candidate TM helix. It is interesting to note that its TM candidacy is weaker than those of Shak-B(44.4) and Ogre, consistent with the hypothesis that this region in Shak-B(42.9) is indeed a cleaved N-terminal signal peptide, and not a signal anchor.

10 of the 18 candidate TM domains have thus been considered. Of the remainder, region A of Unc-7 seems a strong candidate, both because of its acceptance by the Eisenberg, Engleman and *Top_Pred* techniques and because of its homology with the A regions of Shak-B(44.4) and Ogre. Seven potential transmembrane helices are yet to be considered: three B regions (that of Unc-7 having been accepted) and all four C regions. The B regions of Shak-B(44.4), Shak-B(42.9) and Ogre fall into something of a grey area, being accepted by methods with a tendency towards overprediction (see tables 5.1 and 5.3), but rejected by those techniques inclined to underpredict. Thus, in the most conservative interpretation, these regions would be rejected. However, these B regions are all rather confidently accepted by *Top_Pred*, giving scores of 1.40, 1.32 and 1.41, respectively. These scores are higher than those of any of the spurious, overpredicted helices in the proteins considered in table 5.1, or in the larger test set considered by von Heijne (von Heijne, 1992). These scores are also higher than the scores of 2 of the 7 *bona fide* transmembrane helices of Bacteriorhodopsin (table 5.1), with two more Bacteriorhodopsin helices giving scores in the same range (1.33 and 1.34). The B regions of Shak-B(42.9) and Ogre are also accepted by the method of Engleman. Furthermore, given their homology with region B of Unc-7, an apparently irrefutable transmembrane

domain (table 5.3), the B regions of Shak-B(44.4), Shak-B(42.9) and Ogre seem, on balance, more likely than not, to represent true transmembrane domains.

The four remaining candidate helices are the homologous 'C regions' and the status of these is perhaps hardest of all to assess. Because the high homology among these proteins strongly suggests identical topologies, it is tempting to either accept or reject the C region in every case. The *Top_Pred* scores for these regions (0.75-0.89) are in the middle of the putative range. The C regions of Shak-B and Ogre are corroborated (but only just) by the method of Klein *et al*, while the Unc-7 C region is narrowly rejected by this technique, but accepted by the method of Engleman *et al* (see §5.2.2 and §5.2.3). The *Top_Pred* score for the Shak-B region C (when 5.0 kcal/mol charge pair stabilisation energy is granted) is 0.89, close to the 0.90 value observed for the *bona fide* third transmembrane helix of bacteriorhodopsin. On the other hand one overpredicted helix (plus one signal peptide) in table 5.1 generate higher *Top_Pred* scores.

It is very difficult to decide, on theoretical grounds, whether to accept or reject the C region candidate helices. We are therefore left with two best, educated guesses as to the transmembrane topologies of Shak-B(44.4) and its homologues: a three transmembrane model (3TM), which rejects the C regions, and a four transmembrane model (4TM) in which these regions are accepted. It should be stressed once again that these are merely guesses, and are intended to serve as useful hypotheses to be tested in future biochemical analyses. Their derivation has relied wholly on theoretical algorithms for transmembrane helix prediction, even the best of which generate some erroneous predictions (table 5.1). Furthermore, I have relied upon the assertion that the Shak-B(44.4), Shak-B(42.9), Ogre and Unc-7 proteins will, with the possible exception of the potential cleaved signal peptide in Shak-B(42.9) (see §5.4.1), have identical transmembrane topologies. Having said all this, these models are derived from critical application of the best available transmembrane helix and topology predictions. In both 3TM and 4TM models, each proposed helix is either very confidently identified by the *Top_Pred* algorithm (Sipos and von Heijne, 1993) or generates a 'possible' score with this technique and is corroborated by one or more alternative techniques (table 5.3). The 4TM model resembles the suggestions made by the authors of the Unc-7 sequence (Starich, *et al.*, 1993) that up to 4 transmembrane domains are present in that protein. The positions of the proposed helices match precisely. The authors of the Ogre and Shak-

B(42.9) sequences predicted only a single TM domain, corresponding to region D, for each. I believe the 3TM or 4TM model is more likely to reflect the true situation.

It should also be stressed that the term 'extracytoplasmic' has deliberately been used in place of 'extracellular'. Low resolution immunocytochemistry (Watanabe and Kankel, 1992) demonstrates that the Ogre product is localised mainly intracellulary, rather than being expressed on the cell membrane. Whether Ogre protein is in the cytosol or in cytoplasmic organelles has yet to be resolved. However, preliminary anti-Shak-B antibody staining data tentatively suggest cell membrane localisation of at least one form of Shaking-B protein (Marian Wilkin, unpublished results).

Using identical arguments and prediction methods, the membrane topologies of the other Shak-B products proposed in chapters 3 and 4 may now be considered.

5.4.3 Shak-B(14.1)

This proposed product, encoded by the KE2 cDNA (§3.5) is 122 residues in length, the first 120 of which are identical to the first 120 residues of Shak-B(44.4). Hence the first of the proposed membrane spanning domains of Shak-B(44.4) is wholly contained within Shak-B(14.1). Only a fraction of the second proposed TM domain of Shak-B(44.4) is present in Shak-B(14.1), thus this smaller protein is predicted to have only a single TM helix. As in Shak-B(44.4), the topology of this signal anchor domain is predicted to be N terminus cytoplasmic, C terminus extracytoplasmic.

5.4.4 Shak-B(14.0)

This implied product of the P1 cDNA is almost identical to Shak-B(14.1), except that Shak-B(14.0) is one residue shorter and its last residue differs from the second last of Shak-B(14.1). Its anticipated transmembrane topology is identical to that of Shak-B(14.1).

5.4.5 Shak-B(23.3)

This proposed product, encoded by the SIPC726 cDNA is predicted to contain two TM helices, identical in sequence and position to regions A and B of Shak-B(44.4). An alternative splice commitment occurs within the region encoding region C of Shak-B(44.4). The effect of this is to truncate and alter this candidate TM domain in Shak-B(23.3) such that it is too short to span a membrane. Thus Shak-B(23.3) is predicted to have two transmembrane domains, and intracellular N and C termini.

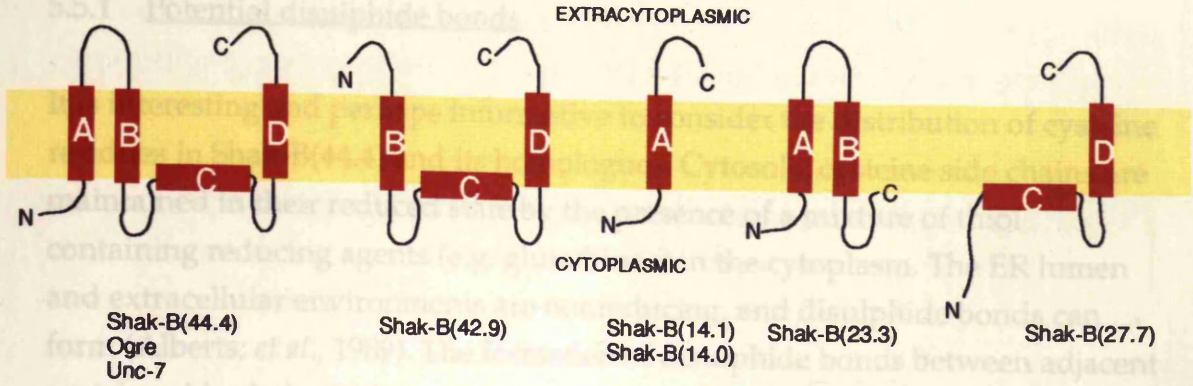
5.4.6 Shak-B(27.7)

Shak-B(27.7), the proposed product of N52, contains regions C and D and thus is expected to have one or two transmembrane domains. The charge difference rule predicts that the N terminus of Shak-B(27.7) will be cytoplasmic, whether region C or region D is the signal anchor domain. Thus this protein species is predicted to have a cytoplasmic N terminus, one or two transmembrane helices and either an extracytoplasmic C terminus or an extracytoplasmic loop and a cytoplasmic C terminus. In other words, in either the 3TM or 4TM models, the predicted topology mimics that of the C terminus of Shak-B(44.4).

These best guess topology predictions for Shak-B proteins and their homologues are shown schematically in Figure 5.7.

Some time after completion of the analysis presented here, a four transmembrane domain model similar to the 4TM model above was proposed for Ogre, Shak-B(neural), Shak-B(lethal) and Unc-7 (Barnes, 1994). Barnes suggested that the protein family be known as OPUS (Ogre, Passover, Unc-7, Shak-B), a name which is perhaps unhelpful, as it suggests Passover and Shaking-B to be separate genes. It is, however, an undeniably catchy acronym, and is therefore almost bound to persist.

5.7a: 3TM model



5.7b: 4TM model

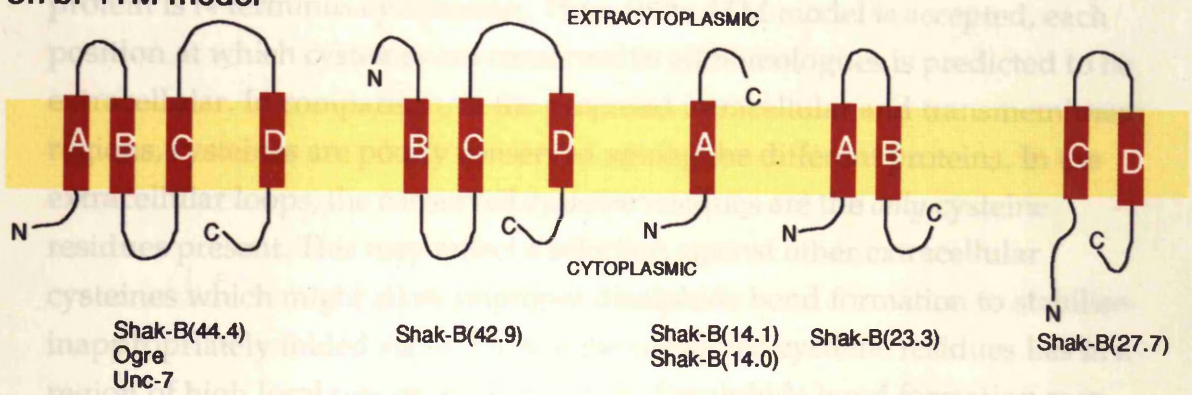


Figure 5.7: Schematic diagram showing transmembrane topology models for Shak-B proteins and their homologues, according to the 3TM and 4TM models. Red rectangles represent potential transmembrane helix domains. Black loops denote polar, non membrane-spanning regions. For details of the derivation of these models, see text.

5.5 THE SEARCH FOR PEPTIDE MOTIFS OF KNOWN FUNCTION

5.5.1 Potential disulphide bonds

It is interesting and perhaps informative to consider the distribution of cysteine residues in Shak-B(44.4) and its homologues. Cytosolic cysteine side chains are maintained in their reduced state by the presence of a mixture of thiol containing reducing agents (e.g. glutathione) in the cytoplasm. The ER lumen and extracellular environments are nonreducing, and disulphide bonds can form (Alberts, *et al.*, 1989). The formation of disulphide bonds between adjacent cysteine side chains reduces the free energy of the protein molecule and is thus favourable. Consider the distribution of cysteine residues in Figure 5.6. The homologous proteins show complete conservation of cysteines at four positions, shown in cyan text. Recall that the most likely topology of each protein is N terminus cytoplasmic. Thus, if the 4TM model is accepted, each position at which cysteines are conserved in all homologues is predicted to be extracellular. In comparison, in the proposed intracellular and transmembrane regions, cysteines are poorly conserved among the different proteins. In the extracellular loops, the conserved cysteine residues are the *only* cysteine residues present. This may reflect a selection against other extracellular cysteines which might allow improper disulphide bond formation to stabilise inappropriately folded states. Each of the conserved cysteine residues lies in a region of high local sequence conservation. Disulphide bond formation may therefore bring together homologous sequences to produce regions of highly conserved tertiary structure.

No such correlation between conserved cysteine residues and extracytoplasmic localisation is seen if the 3TM model is accepted. However, the pattern of cysteine residues is only an observation. It cannot in itself be taken as corroborating evidence for the 4TM model.

5.5.2 Searching for peptide motifs represented in the PROSITE database

The MacPattern program (Fuchs, 1991) was used to search Shak-B proteins and their homologues for other protein motifs contained within the PROSITE (Bairoch, 1991) database. Release 12.0 (June, 1994) of this database was obtained by anonymous ftp to the EMBL fileserver (ftp.EMBL-heidelberg.de). MacPattern is an elegant and powerful program for finding recognisable

peptide motifs, however many peptide motifs which can act as e.g. phosphorylation sites, are short (perhaps depending critically on two residues only) and degenerate (e.g. one of the residues must be basic). The false positive rate resulting from searches with such short peptide motifs is therefore high.

Given this caveat, I have included in the appendix to this chapter a series of tables showing the positions of potentially functional peptide motifs in Shak-B(44.4) and its homologues. Most of the motifs considered can only represent functional sites if they are cytoplasmically located. For N-linked glycosylation sites, the opposite is true. Where the predicted cytoplasmic or extracytoplasmic location of a motif is inappropriate to its possible function in one or both of the topology models favoured, this is noted. However the potential sites found are not well conserved among the homologous proteins, and, given the high rate of false positives in the detection of these short motifs, their positions cannot be used to assess alternative topology models in the absence of biochemical evidence verifying the function of any of the identified sites.

The sequences of Shak-B proteins and their homologues were also used for searches of the current Blocks database (Wallace and Henikoff, 1992). The searches were done remotely by E-mail to the Blocks server at blocks@howard.fhcrc.org. No entries in the Blocks database release 8.0 showed significant homology to Shak-B proteins or their homologues.

APPENDIX TO CHAPTER 5

5A.1: The *Signify* program

Included here is the complete, annotated code for the *Signify 1.1* program, written for Macintosh in the THINK-C programming language (Symantec). The program analyses eukaryotic protein sequences to search for cleavable N-terminal signal peptides, according to the method of von Heijne (von Heijne, 1986), and outputs the results both to screen and printer.

5A.1 Full, annotated code for the *Signify* program

In the listing that follows the text in Helvetica font is actually included in the program while that in `Courier font` is commentary and is not part of the program listing.

```
/* Signify */

#define PRINT_RETURN printf("\n");
#define PRINT_TAB printf("\t");
#define MAX_PEP 1000
```

```
Initial #define commands. Print return and Print tab commands are
simply to make the code more user friendly. The #define MAX_PEP 1000
command sets a maximum limit to the size of the peptide sequence that
may be input. To analyse very large coding sequences, this line alone
should be changed, and a greater number substituted.
```

```
#include <stdlib.h>
#include <stdio.h>
#include <math.h>
#include <console.h>
#include <time.h>
```

```
#includes of functions required by SIGNIFY 1.1. Most are ANSI standard
library functions, though <console.h> which handles the program's
screen printing, in unique to THINK-C.
```

```
/* initiate global variables */

FILE    *consolePtr;
/*                                                  Pointer to console window */

float   counts[25][16];
/*      floating point array in which to input Von Heijne's data
        pep[MAX_PEP], response[16], ver[MAX_PEP], name [50], printout[100];
                                                declare arrays for peptide strings    */

char    pep[MAX_PEP], response[16], ver[MAX_PEP], name [50], printout[100];
/*                                              declare arrays for peptide strings    */

int             i,j, k, length, m, n, w, adjustor;
int             threshold = -100;
/*                                                  fire up some integer variables    */

long double alg;
```

```
time_t now;
struct tm *date;                        /* get ready to tell time */
char s[80];

/* Function code starts here */

printtime()

{
time (&now);
date = localtime(&now);
strftime (s,80, "%c", date);
sprintf (printout,"%s\n\n",s);
hardcopy();
}
This routine gathers time and date information which is echoed to the
printer using the hardcopy() function.

void flush ()

{
        while (getchar()!='\n')
        ;
}
Clears the input buffer of any unwanted remnants of responses to old
prompts.

purge ()

{
int             i,j;
for             (j=0; j<16; j++)
                {
                        for     (i=0; i<25; i++)
                                {
                                counts[i][j] = 0;
                                }
                        response [j]=0;
                }
}
Sets zero values of the floating point array into which the data of
von Heijne (von Heijne, 1986) will be written.

setprinter()
{
        cecho2printer (consolePtr);             /* Echo screen output to printer */
        sprintf (printout,"\nSIGNIFY of %s\t\t\t",name);
        hardcopy();
        printtime();
        sprintf(printout,"\n\nThreshold is %d\n\n",threshold);
        hardcopy();
}
This routine writes the title on the printed output.

setthreshold()

{
while (threshold==-100)
        {
                printf("Please input a threshold above which to show windows\t");
                scanf("%d",&threshold);
                flush();
```

```
        }
}
```

Prompts for a value of a threshold above which to show potentially significant windows. The while loop repeats the prompt until a meaningful answer is provided. Von Heijne (1986) showed that when the N terminal 40 amino acids of a variety of proteins is scanned, a value of 3.5 gives a highly selective distinction between cytosolic and membrane-bound sequences.

optioncontrol()

This function allows the choice between a manually input sequence and one read from a disk file. I am not a computer expert, and personal coding limitations mean that the sequence source file must be in the same directory as SIGNIFY, and amino acids must be the only letters in the sequence files.

```
{
        char response = 0;
        printf("SIGNIFY supports EITHER keyboard entry of small (<60 residues)
\nsequences");
        printf(" OR input from a sequence file on disk.\n");
        while ((response!='D')&&(response!='K'))
        {
                printf("\n\nPlease specify keyboard (K) or disk file (D) input\t");
                scanf("%s",&response);
                if (response>'Z') response=response-32;
                flush();
        }
```

While loop demands a response of D or K. The line if (response>'Z') response=response-32; supports lower case entry of the decision letter.

```
        switch (response)
        {
        case 'K':
                        input();
                        break;
        case 'D':
                        readafile();
                        break;
        }
```

switch is definitely my favourite C command. This switch loop allows the program to flow to the keyboard input function input() or the disk file input function readafile() according to the response to optioncontrol().

```
}


input()
{
        char    *pepAddress, *verAddress;

        PRINT_RETURN

                for (i=0;i>=499;i++)
                {
                        pep[i] = 0;
                        ver[i] = 0;                                     /*      clear arrays for string
                }
comparisons     */

        printf("Name of peptide sequence: ");
        scanf("%s",name);
```

```
        flush();

        printf("\n\nPlease input peptide sequence (large case ):\n\n");
        scanf ("%s", pep);
        flush();

        printf("Please verify:\n\n");
        scanf ("%s", ver);
        flush();
```

The input() routine thus far has asked for the name and contents of a sequence, and a second inputting of the sequence to verify that it is correct.

```
        pepAddress = &pep[0];
        verAddress = &ver[0];
```

Need to set some address variables in order to use the strcmp() function to compare the intial sequence and its verification:

```
        /* establish string addresses to allow string comparisons          */

        while (strncmp(pepAddress, verAddress))
                {
                                                /*      while loop for verification failure */


                        putc (7, consolePtr);           /* BEEP */
                        printf ("verification failed: please try again\n\n");
                        scanf ("%s", pep);

                        printf ("please verify new attempt\n\n");
                        scanf ("%s", ver);

                }
```

The strcmp() function returns a nonzero value only if the initial peptide is different in sequence to its verification. If this is the case, both peptide and verification sequences are prompted for again. This while loop goes on forever until the user gets the two sequences the same.

\

```
        length = strlen (pep);
        printf("\nYou have input a peptide sequence of %d residues\n\n",length);

                if(length<14)
                                {
                                 printf ("sequence is too short to analyse\n");
                                finishoff();
                                exit(0);
                                }
```

The program closes down politely if the input sequence is too short to analyse. Von Heijne's weight matrix includes from the -13 to the +2 positions relative to the potential cleavage site under investigation. This program is not equipped to deal with any sequences shorter than one full window's worth. The finishoff() routine tells the user how many significant windows have been found, closes down the console window, and turns off the printer output.

```
                else
                                PRINT_RETURN
}
```

```
readafile ()
The routine for disk file input.
{
FILE    *fp;            initiates a pointer to the disk file
int     b=0;
int     c=0;
adjustor = -1;          the adjustor variable is used later to subtract the
                        EOF character when declaring peptide length
printf("\nI can only interpret text only, upper case, peptide sequences\n");
printf ("in the same folder as the SIGNIFY application.\n\n");
printf("Please tell me the name of your peptide to analyse... ");
scanf("%s", name);

fp=fopen(name,"r");                             /* Opening Disk File */
if      (fp != NULL )                           function exits if the file can't
                                                be found

        {while ((c=fgetc(fp))!=EOF)             function gives up once End Of
                                                File is reached
        {
        pep[b]=c;                               writes file into peptide array
        b++;
        if (b>(MAX_PEP-1)) break; /* sets peptide size limit of arrays */
        }
        fclose(fp);                                     /* Closing Disk File */
        }
    else {
        printf("\nI'm sorry, I can't find this sequence.");
        putc (7, consolePtr);     /*BEEP*/
        fclose (consolePtr);
        exit(0);
        }
}
There follows a laborious function which initialises the variables for
the weight matrix array. Zero values are not set again, but remain
zero after the purge() function

vonheijne ()

    {

/*      Defining values for occurrence of alanine       */
counts[0][0] = 14.5;    counts[0][1] = 16;              counts[0][2] = 13;
counts[0][3] = 14;              counts[0][4] = 15;              counts[0][5] = 20;
counts[0][6] = 18;              counts[0][7] = 18;              counts[0][8] = 17;
counts[0][9] = 25;              counts[0][10] =15;              counts[0][11]= 47;
counts[0][12]= 6;               counts[0][13] =80;              counts[0][14]= 18;
counts[0][15]= 6;

/*      Defining values for occurrence of cysteine      */
counts[2][0] = 4.5;             counts[2][1] = 3;               counts[2][2] = 6;
counts[2][3] = 9;               counts[2][4] = 7;               counts[2][5] = 9;
counts[2][9] =14;               counts[2][7] = 6;               counts[2][8] = 8;
counts[2][9] = 5;               counts[2][10] = 6;              counts[2][11]= 19;
counts[2][12] = 3;              counts[2][13] = 9;              counts[2][14] = 8;
counts[2][15] = 3;

/* Defining values for the occurrence of aspartate      */
counts[3][0] = 8.9;
counts[3][9] = 5;               counts[3][10]= 3;
counts[3][12] =5;               counts[3][14] =10;
counts[3][15] = 11;
```

```
/*      Defining values for the occurrence of glutamate  */
counts[4][0] =10.0;
counts[4][4] = 1;
counts[4][9] = 3;                counts[4][10]= 7;
counts[4][12] =7;                counts[4][14] =13;
counts[4][15] =14;

/*      Defining values for the occurrence of phenylalanine */
counts[5][0] =5.6;               counts[5][1] = 13;              counts[5][2] = 9;
counts[5][3] = 11;               counts[5][4] = 11;              counts[5][5] = 6;
counts[5][6] = 7;                counts[5][7] = 18;              counts[5][8] = 13;
counts[5][9] = 4;                counts[5][10]= 5;
counts[5][12] =13;               counts[5][14] =6;
counts[5][15] =4;

/*      Defining values for the occurrence of glycine      */
counts[6][0] =12.1;              counts[6][1] = 4;               counts[6][2] = 4;
counts[6][3] = 3;                counts[6][4] = 6;               counts[6][5] = 3;
counts[6][6] = 13;               counts[6][7] = 3;               counts[6][8] = 2;
counts[6][9] = 19;               counts[6][10]= 34;              counts[6][11]= 5;
counts[6][12] =7;                counts[6][13] =39;              counts[6][14] =10;
counts[6][15] =7;

/*      Defining values for the occurrence of histidine */
counts[7][0] =3.4;
counts[7][6] = 1;                counts[7][7] = 1;
counts[7][9] = 5;
counts[7][12] =6;                counts[7][14] =4;
counts[7][15] =2;

/*      Defining values for the occurrence of  isoleucine  */
counts[8][0] =7.4;               counts[8][1] = 15;              counts[8][2] = 15;
counts[8][3] = 8;                counts[8][4] = 6;               counts[8][5] = 11;
counts[8][6] = 5;                counts[8][7] = 4;               counts[8][8] = 8;
counts[8][9] = 5;                counts[8][10]= 1;               counts[8][11]= 10;
counts[8][12] =5;                counts[8][14] = 8;
counts[8][15] =7;

/*      Defining values for the occurrence of  lysine */
counts[10][0] =11.3;
counts[10][4] = 1;
counts[10][7] = 1;
counts[10][10]= 4;
counts[10][12] =2;               counts[10][14] =11;
counts[10][15] =9;

/*      Defining values for the occurrence of  leucine */
counts[11][0] =12.1;     counts[11][1] = 71;             counts[11][2] =68 ;
counts[11][3] =72 ;              counts[11][4] = 79;             counts[11][5] = 78;
counts[11][6] = 45;              counts[11][7] = 64;             counts[11][8] = 49;
counts[11][9] = 10;              counts[11][10]= 23;             counts[11][11]= 8;
counts[11][12] =20;              counts[11][13] =1;              counts[11][14] =8;
counts[11][15] =4;

/*      Defining values for the occurrence of  methionine */
counts[12][0] =2.7;              counts[12][2] = 3;
counts[12][3] = 7;               counts[12][4] = 4;              counts[12][5] = 1;
counts[12][6] = 6;               counts[12][7] = 2;              counts[12][8] = 2;
counts[12][12] =1;               counts[12][14] =1;
counts[12][15] =2;

/*      Defining values for the occurrence of asparagine */
```

```
counts[13][0] =7.1;          counts[13][2] = 1;
counts[13][4] = 1;           counts[13][5] = 1;
counts[13][9] = 3;           counts[13][10]= 3;
counts[13][12] =10;          counts[13][14] =4;
counts[13][15] =7;

/*    Defining values for the occurrence of proline */
counts[15][0] =7.4;          counts[15][1] = 2;
counts[15][3] = 2;
counts[15][6] = 4;           counts[15][7] = 1;          counts[15][8] = 8;
counts[15][9] = 20;          counts[15][10]= 14;
counts[15][12] =1;           counts[15][13] =3;
counts[15][15] =22;

/*    Defining values for the occurrence of  glutamine */
counts[16][0] =6.3;
counts[16][4] = 1;
counts[16][6] =6 ;           counts[16][7] = 1;
counts[16][9] = 10;          counts[16][10]= 8;
counts[16][12] =18;          counts[16][13] =3;          counts[16][14] =19;
counts[16][15] =10;

/*    Defining values for the occurrence of  arginine */
counts[17][0] =7.6;          counts[17][1] = 2;
counts[17][6] = 1;
counts[17][9] =7 ;           counts[17][10]= 4;
counts[17][12] =15;          counts[17][14] =12;
counts[17][15] =9;

/*    Defining values for the occurrence of  serine */
counts[18][0] =11.4;   counts[18][1] = 9;          counts[18][2] = 3;
counts[18][3] = 8;           counts[18][4] = 6;          counts[18][5] = 13;
counts[18][6] = 10;          counts[18][7] = 15;         counts[18][8] = 16;
counts[18][9] = 26;          counts[18][10]= 11;         counts[18][11]= 23;
counts[18][12] =17;          counts[18][13] =20;         counts[18][14] =15;
counts[18][15] =10;

/*    Defining values for the occurrence of  threonine */
counts[19][0] =9.7;          counts[19][1] = 2;          counts[19][2] = 10;
counts[19][3] = 5;           counts[19][4] = 4;          counts[19][5] = 5;
counts[19][6] = 13;          counts[19][7] = 7;          counts[19][8] =7 ;
counts[19][9] = 12;          counts[19][10]=6 ;          counts[19][11]= 17;
counts[19][12] =8;           counts[19][13] =6;          counts[19][14] =3;
counts[19][15] =10;

/*    Defining values for the occurrence of valine */
counts[21][0] =11.1;   counts[21][1] = 20;         counts[21][2] = 25;
counts[21][3] = 15;          counts[21][4] = 18;         counts[21][5] = 13;
counts[21][6] = 15;          counts[21][7] = 11;         counts[21][8] = 27;
counts[21][10]= 12;          counts[21][11]= 32;
counts[21][12] =3;           counts[21][14] =8;
counts[21][15] =17;

/*    Defining values for the occurrence of tryptophan */
counts[22][0] =1.8;          counts[22][1] = 4;          counts[22][2] = 3;
counts[22][3] =3 ;           counts[22][4] = 1;          counts[22][5] = 1;
counts[22][6] = 2;           counts[22][7] = 6;          counts[22][8] = 3;
counts[22][9] = 1;           counts[22][10]= 3;
counts[22][12] =9;           counts[22][14] =2;

/*    Defining values for the occurrence of tyrosine */
counts[24][0] =5.6;          counts[24][2] = 1;
```

```
counts[24][3] = 4;
counts[24][6] = 1;          counts[24][7] = 3;          counts[24][8] = 1;
counts[24][9] = 1;          counts[24][10]= 2;
counts[24][12] =5;          counts[24][14] =1;
counts[24][15] =7;

printf ( "******  SIGNIFY ******\n\n");

As a bit of a celebration after all that information, the function
writes the title of the program to the console
}

weightmatrix ()
Here the calculations start. Weightmatrix() invokes calculate() to
determine score of current window. If that score is greater than
threshold, the position and sequence and score of the window are
printed on the console and the printer
{
        length = strlen (pep) - adjustor;
        printf ("\nPeptide length is %d\n\n", length);
        sprintf (printout,"\nPeptide length is %d\n\n", length);
        hardcopy();
        m=0;    j=0;

        while (m<=(strlen(pep)-16))          /*     m loop recurs for each window */
                {

                        calculate();
                        k=15;
                        if (alg>threshold)              /* only print >threshold windows */
                        {
                                w++;
                                putc (10,consolePtr);
                                printf(" position %d\t",m);
                                sprintf(printout," position %d,\twindow:\t",m);
                                hardcopy();
                                printf ("window:\t ");

                                i=1; k=15;
                                while ((i-1)<k)         /* n is position within window */
                                {

                                checkchar();


                                printf("%c",pep[m+i-1]);
                                putc(pep[m+i-1], consolePtr);
                                i++;

                                                                /* window printed */

                                }

                        putc (9,consolePtr);            /* Printer Tab */
                        PRINT_TAB                               /* Console Tab */

                        printf ("\tscore : %G",alg);
                        sprintf (printout, "\tscore: %G\t",alg);
                        PRINT_RETURN
                        hardcopy();

                        for (i=0;i<alg;i++)             /* Print dots on printer */
```

```
                printout[i]='.';

                printout[i]=0;          /* Terminate string before Printing it */
                hardcopy();
                }

        m++; n=0;
    }

}

calculate   ()      calculate() calculates the score for the current
                    window

{
    i=0;    j=0;
    alg=0;
    k=15;
    while (i<k)
    {

        i++;

        j++;    /* position in window counter not moved by checkchar */

        checkchar();


        /* If array value for amino acid is non zero, add ln of score,
                and subtract ln of its expected accurrence
    */

        if (counts[pep[m+i-1]-'A'][j]>0)
            {
                alg = alg+log(counts[(pep[m+i-1]-'A')][j]);
                alg = alg-log(counts[(pep[m+i-1]-'A')][0]);
            }


        /* The else condition deals with zero values of array */

        else
            {
                switch (j)
                {
                case 11:
                case 13:
                    alg=alg-log(161);
                    break;

                default:
                    alg=alg-log(counts[pep[m+i-1]-'A'][0]);
                    break;
                }
            }
    } /* i position counter loops back here */
}

hardcopy()              /* This routine prints out printer buffer in
                                printout [100] character array      */
```

```
{
        int q=0;
        while (printout[q]!=0)
                {
                        putc(printout[q], consolePtr);
                        q++;
                }
}
checkchar()
This function allows skipping of non amino acid characters, by
incrementing both the position within the window being scanned and the
size of the window if a nonsense character is found. It also allows
for lower case sequences. The function often refers back to itself
(see flow diagram (Figure 5.2)). If one character is nonsense the
function invokes itself to check the next until a meaningful character
is reached

{               if(pep[m+i-1]==0)       this would imply the end of the
                                                                peptide
                {
                        finishoff();
                        exit(0);
                }

        /* entertain lower case entries */
                while ((pep[m+i-1]>'a')&&(pep[m+i-1]<'z'))
                {
                        pep[m+i-1]=(pep[m+i-1]-32);
                }


        /* move on if you find non AA characters */
                while ((pep[m+i-1]<'A')||(pep[m+i-1]>'Y'))
                {
                        i++;
                        k++;
                        checkchar();
                }

        while ((pep[m+i-1]=='B')||(pep[m+i-1]=='J'))
                {
                        i++;
                        k++;
                        checkchar();
                }
        while ((pep[m+i-1]=='O')||(pep[m+i-1]=='U'))
                {
                        i++;
                        k++;
                        checkchar();
                }
        while (pep[m+i-1]=='X')
                {
                        i++;
                        k++;
                        checkchar ();
                }
}

finishoff()
```

```
{
sprintf(printout,"\n\n%d significant windows found.\t",w);

hardcopy();

sprintf(printout,"\n");
hardcopy();                                    /* spits out last of printer buffer */

printf("\n%d significant windows found.\t",w);
printf( "\n\nSignify program complete: press return to exit ");

fclose (consolePtr);
}


main()  the program starts performing tasks from here

{

char title[10]=" SIGNIFY";
char *titlePtr=&title[0];              /*creating Pascal string title*/
title [0]=7;

console_options.title=(unsigned char*)titlePtr;
console_options.nrows=28;

consolePtr=fopenc();
        if (consolePtr==NULL)
        exit(-1);
Declares the title of the console and opens the console window
purge();

w=0;

vonheijne();

optioncontrol();

setthreshold();

setprinter();

weightmatrix();

fputc (10, consolePtr);

finishoff();


}
```

5A.2 Potentially functional peptide motifs contained within Shak-B(44.4) and its homologues

Of the Shak-B protein species proposed, only Shak-B(44.4) and Shak-B(42.9) are included here. Neither the smaller protein forms (e.g. Shak-B(14.1)), nor the Shak-B(42.9) protein in the *shak-B*^{Passover} mutant contain any identifiable peptide sequence motifs which are not present in their larger relatives.

Protein	Sequence	Comments
Shak-B(44.4)	43: TTR	Predicted to be in signal anchor
	76: TLK	Predicted to be extracytoplasmic
	126:TPR	Predicted to be in second TM domain
	296:SPR	Cytoplasmic only in 4TM model
Shak-B(42.9)	31: TTR	Predicted to be in signal peptide or signal anchor
	92: TIK	Predicted to be extracytoplasmic
	115:TPR	Predicted to be in second TM domain
	285:SPR	Cytoplasmic only in 4TM model
Ogre	7: SLK	
	125:TPK	Predicted to be in second TM domain
	306:SNR	Cytoplasmic only in 4TM model
	319:SRK	Cytoplasmic only in 4TM model
Unc-7	43: SKK	
	162:SAK	Predicted to be in signal anchor
	211:SRR	Predicted to be extracytoplasmic
	461:TLR	Cytoplasmic only in 4TM model

Table 5.A.1: Potential protein kinase C phosphorylation sites
Consensus: [S/T]-X-[R/K]; phosphorylation at S/T
Reference: (Woodget, *et al.*, 1986)

Protein	Sequence	Comments
Shak-B(44.4)	97: SDGD	Predicted to be extracytoplasmic
	152:SEAE	
	220:TDQE	Cytoplasmic only in 3TM model

	246:SSGE	Cytoplasmic only in 3TM model
Shak-B(42.9)	64: TVVD	Predicted to be extracytoplasmic
	141:SEAE	
	209:TDQE	Cytoplasmic only in 3TM model
	235:SSGE	Cytoplasmic only in 3TM model
Ogre	75: TMPD	Predicted to be extracytoplasmic
	151:TREE	
	356:SKHD	Cytoplasmic only in 4TM model
Unc-7	6: SNPE	
	67: TPLE	
	178:TFTD	Predicted to be extracytoplasmic
	289:TNID	
	375:TLCD	Cytoplasmic only in 3TM model

Table 5.A.2: Potential casein kinase II phosphorylation sites
Consensus:[S/T]-X-X-[D/E]; phosphorylation at S/T
Reference:(Pinna, 1990)

Protein	Sequence	Comments
Shak-B(42.9)	17: NATV	Predicted to be in signal peptide or signal anchor
Ogre	146:NITI	Not predicted to be extracytoplasmic
	304:NASN	Not predicted to be extracytoplasmic

Table 5.A.3: Potential N-linked glycosylation sites
Consensus: N-[not P]-[S/T]
Reference: (Miletich and Broze, 1990)

Protein	Sequence	Comments
Ogre	RNLDPVIY	Cytoplasmic only in 4TM model

Table 5.A.4: Potential tyrosine phosphorylation site
Consensus: [R/K]-x(2 or 3)-[D/E]-x(2 or 3)-Y
Reference: (Hunter, 1982)

Protein	Sequence	Comments
Unc-7	463:RKFT	Cytoplasmic only in 4TM model

Table 5.A.5: Potential cAMP dependent protein kinase phosphorylation site

Consensus:[R/K](2)-x-[S/T] (phosphorylation at S or T)

Reference: (Glass, *et al.*, 1986)

Six

Discussion

6.1 ON THE STRUCTURE OF THE *SHAKING-B* LOCUS

In chapters 3 and 4, I have presented details of the structures of seven overlapping cDNAs: KE2(1.8), P1, SIPC8, SIPC737, SIPC726, N52[1], and SIPC224. Each of these represents a transcript derived from the 19E3 region of the *Drosophila melanogaster* X chromosome, the area where the *shaking-B* locus is known to reside.

The largest open reading frame (ORF) contained within these cDNA clones is the 372 codon ORF of SIPC8. I have shown that three chromosomes bearing lethal mutations of *shaking-B* exhibit lesions which disrupt this coding sequence, providing firm evidence both that these seven overlapping cDNAs are derived from the *shaking-B* locus, and that the 372 amino acid ORF is translated to make an essential Shaking-B protein, here referred to as Shak-B(44.4).

Two of the lethal *shaking-B* lesions found are identical: both *shak-B*$^{R\text{-}9\text{-}29}$ and *shak-B*^{EC201} alleles show a G»A transition at base 1656 of the SIPC8 sequence, which changes the TGG (tryptophan) codon 273 to a TGA (stop) codon. Both of these alleles complement neither lethal nor neural alleles of *shaking-B*, implying that they disrupt both genetic functions at this locus, and suggesting that essential and neural Shaking-B proteins share the region of reading frame disrupted by the *shak-B*$^{R\text{-}9\text{-}29}$/*shak-B*^{EC201} lesion. Conversely, the *shak-B*^{L41} allele is lethal but fully complements neural *shaking-B* alleles, thus must disrupt a region required for the essential, but not for the neural function. The *shak-B*^{L41} lesion is a 17 base pair deletion, removing the start codon which initiates the Shak-B(44.4) protein, implying that Shak-B(neural) protein(s) must be translated from a distinct start site. The P2.4 cDNA presented by Krishnan and coworkers (Krishnan, *et al.*, 1993) represents another alternatively processed transcript related to those presented here. Its open reading frame begins at a start codon upstream of the translation start disrupted by the *shak-B*^{L41} lesion, while downstream its coding sequence is shared with that of SIPC8 such that it too is disrupted by the *shak-B*$^{R\text{-}9\text{-}29}$ and *shak-B*^{EC201} mutations, making it a

---

1 Please note, once again, that the sequence of the N52 cDNA was derived by Shuqing Ji.

strong candidate for a neural transcript of *shaking-B*. This candidacy is greatly strengthened by the isolation, by Krishnan and colleagues, of mutations associated with the neural alleles *shaking-B*^{2} and *shaking-B*^{Passover} in the upstream coding sequence unique to the P2.4 splice form. Taken together, these data suggest a simple molecular model to account for the complex complementation relationships at *shaking-B* (Figure 4.10): Shak-B(44.4) is an essential Shak-B protein, while Shak-B(42.9), the product of P2.4, is required for adult nervous system development. *shaking-B* mutations that are lethal (non neural) and neural (non lethal) map to the unique upstream coding regions of SIPC8 and P2.4 respectively, while lesions causing mutations which are lethal *and* neural map to the common downstream coding sequences shared by the ORFs of these two transcripts.

This model is fine, as far as it goes, but it is based upon the positions of only five of the eight *shaking-B* mutations and the structures of only two of the eight cDNA sequences discussed thus far. Two questions therefore arise: where are the other *shak-B* mutations and what, if anything, do transcripts represented by the remaining cDNAs do? These vexed questions will be addressed in turn.

6.1.1 The positions of the remaining *shaking-B* mutations

In chapters 3 and 4, I described experiments aimed at detecting the molecular lesions underlying nine *shaking-B* alleles: *shak-B*^{2}, *shak-B*^{Passover}, *shak-B*^{L41}, *shak-B*^{EF535}, *shak-B*^{EC201}, *shak-B*$^{R-9-29}$, *shak-B*^{E81}, *shak-B*$^{17-360}$, and *shak-B*^{HM437}. Of these nine, I have accounted for four (*shak-B*^{L41}, *shak-B*^{EF535}, *shak-B*^{EC201}, and *shak-B*$^{R-9-29}$). Krishnan *et al.*, (1993) have described the molecular bases of the *shak-B*^{Passover} and *shak-B*^{2} alleles. *shak-B*^{Passover} is due to a C»T transition replacing a CGG (arginine) codon with a TGG (tryptophan), at position 33 of the Shak-B(42.9) protein (§5.4.1), while *shak-B*^{2} is caused by a T»A transversion turning a TTA (leucine) codon into a TAA (stop) at position 22 of the same protein. This same report (Krishnan, *et al.*, 1993) also demonstrated that the P element allele used by Krishnan and colleagues to clone *shaking-B* was caused by P element insertion into the first exon of P2.4, in its upstream, untranslated region. This mutation causes the hypomorphic *shak-B(neural)* allele *shak-B*^{njP181}, whose position and genetic nature fit neatly into the molecular model of *shak-B* complementation (Figure 4.10) which was recapitulated above.

Thus Krishnan and colleagues have described the molecular natures of all three *shak-B* alleles which affect only *shak-B(neural)*, while here I have described the lesion underlying the *shak-B*^{L41} allele (*shak-B*^{EF535} being identical (§3.5.5.b)), so accounting for the mutations affecting only the *shak-B(lethal)* function. My search for a further five mutations, affecting both neural and lethal *shak-B* functions was successful on two counts (*shak-B*$^{R-9-29}$ and *shak-B*^{EC201}). Where, then, are the others? Very recently this question has been answered in another publication from the Wyman laboratory (Krishnan, *et al.*, 1995). This paper reports the sequence of a *shak-B(lethal)* cDNA, though the developmental stage from which the clone was derived was not reported. When compared to SIPC8, the new sequence contains no exon A, and a truncated version of exon B. Its coding sequence is, however, identical to that of SIPC8. Krishnan *et al.*, (1995) confirm the positions and natures of the molecular lesions causing the *shak-B*^{L41}, *shak-B*^{EF535}, *shak-B*$^{R-9-29}$ and *shak-B*^{EC201} muations, as reported here. In addition, they describe point mutations underlying *shak-B*^{E81} and *shak-B*^{HM437}. The *shak-B*^{E81} lesion is reported as a T»A transversion changing a TGT (cysteine) codon to a TGA (stop) codon. The mutation is in the codon equivalent to residue 151 of Shak-B(44.4), lying in the exon D coding region common to P2.4 and SIPC8 (Figure 4.8). *shak-B*^{HM437} was reported as a T»A transversion converting TTA (leucine) to TAA (stop) at the position equivalent to amino acid 310 of Shak-B(44.4), at the start of exon H of SIPC8. Once again this coding region is common to the Shak-B(42.9) and Shak-B(44.4) conceptual proteins, thus both *shak-B*^{HM437} and *shak-B*^{E81} fit neatly into the molecular model (Figure 4.10). In addition *shak-B*$^{17-360}$ was declared (Krishnan, *et al.*, 1995) to be a complex rearrangement affecting coding exons common to P2.4 and the new *shak-B(lethal)* cDNA, though no further details were reported. Once again this is consistent with the molecular model.

Why were these latter 3 *shaking-B* mutations not detected here? A number of possibilities present themselves. Firstly, might these polymorphisms have been missed on reading the sequencing gels? For *shak-B*$^{17-360}$ this question is hard to address, as the exact nature of the lesion was not reported. However, reexamination of the other relevant sequences (PCR between PF1 and PF3 (see Figure 4.8), sequenced with the PF3 primer, for *shak-B*^{E81} and PCR between PK6 and PM1, sequenced with the PK8 primer for *shak-B*^{HM437}) reveals clear and emphatic negative results. No sequence polymorphisms are present at the reported sites of these lesions. Instead, single, wild type sequences are clearly

visible. The clear negative result for the *shak-B*^{HM437} lesion is shown (incidentally) on Figure 4.7. Data for *shak-B*^{E81} are not shown.

There are three possibilities for these conflicting results. Either (i) the primers used for my PCR failed to amplify successfully from the mutant chromosomes in question, (ii) incorrect fly stocks were used in these experiments, or (iii) (a formal, but remote possibility) some of the data presented by Krishnan and colleagues are incorrect. In the case of *shak-B*^{HM437}, the first of these possibilities is excluded, as sequence heterozygosity is observed in the sequence derived from the exon H region in my '*shak-B*^{HM437}' stock (Figure 4.5). This suggests that my own negative results may have been due to the wrong stock being used. In the case of *shak-B*^{E81}, no heterozygosity is observed in the region reported to contain the mutant lesion, thus it is also possible that the primer pair used failed to amplify efficiently from the mutant chromosome.

6.1.2 How many *shaking-B* transcript forms are actually functional?

6.1.2.a Proteins required for viability

The fact that all *shak-B(lethal)* alleles fail to complement each other indicates that they all disrupt the same genetic function. In other words, a protein exists whose structure (and/or expression) is disrupted by all of the lesions causing lethal alleles of *shak-B*. Of all the conceptual Shak-B proteins to date extrapolated from cDNA sequences, only the Shak-B(44.4) sequence fulfils this criterion, strongly suggesting that this protein is indeed *required* for viability. There remains, however, the formal caveat that Shak-B(44.4) might not be required for viability, but that another (as yet undetected) protein which also encompasses all of the coding regions disrupted by *shak-B(lethal)* lesions is.

6.1.2.b Proteins sufficient for viability

The appropriate experiment to test this caveat is to create transgenic flies carrying a SIPC8 construct and to test whether the construct can rescue animals deficient for the *shaking-B* locus. If such a procedure resulted in viable, wild-type flies then it would prove that Shak-B(44.4) is *sufficient* for viability. In this event, Shak-B(44.4) is also *required* for viability, except in the unlikely event that another (as yet undetected) protein disrupted by the same lesions and with

complementary biochemical function is also present. If the construct did not effect rescue, despite *in situ* hybridisation revealing wild-type mRNA expression patterns, then it might be because one or more smaller putative Shak-B proteins described here is/are also essential for viability. In this instance, assuming that some but not all putative Shak-B(lethal) proteins are essential, and given that the smaller putative Shak-B proteins are each disrupted by some but not all *shak-B(lethal)* lesions, then the situation might arise wherein some *shak-B(lethal)* alleles are rescued by the SIPC8 transgene while others are not. Thus, for example, if *shak-B*^{E81} but not *shak-B*^{L41} were found to be rescuable by a given SIPC8 transgene insertion, then this might reflect that Shak-B(14.1) and/or Shak-B(14.0) and/or Shak-B(23.3) is/are also required for viability, while Shak-B(27.7) is not. A corollary of this is that it is prudent to test for SIPC8 transgene rescue in heterozygous genotypes, such as *shak-B*^{L41}/*shak-B*^{HM437}, as no known Shak-B protein except Shak-B(44.4) is currently known to be disrupted by both these mutations.

The results of transgenic rescue experiments are eagerly awaited (Martin Todman and Lucy Stebbings, experiments in progress). In the meantime , it is worth considering what (if any) functions the different Shak-B protein species have. In this discussion I will assume, for the sake of argument, that all of the conceptual Shak-B proteins proposed here are genuinely translated. Any individual Shak-B protein may either be essential (as the genetic data strongly suggest for Shak-B(44.4)), non-essential but still functional, or non-functional. In the latter case a protein may either be a 'deliberately' non-functional molecule translated from a non-functional mRNA splice form, reflecting the use of differential splicing to turn gene expression on and off (see, for example, Bell, *et al.*, 1988), or alternatively it may be 'accidental': a non-functional but harmless protein (e.g. Erickson, 1993). In this regard it is interesting to consider the codon usage within the Shak-B(44.4) ORF. As shown in Figure 6.1, below.

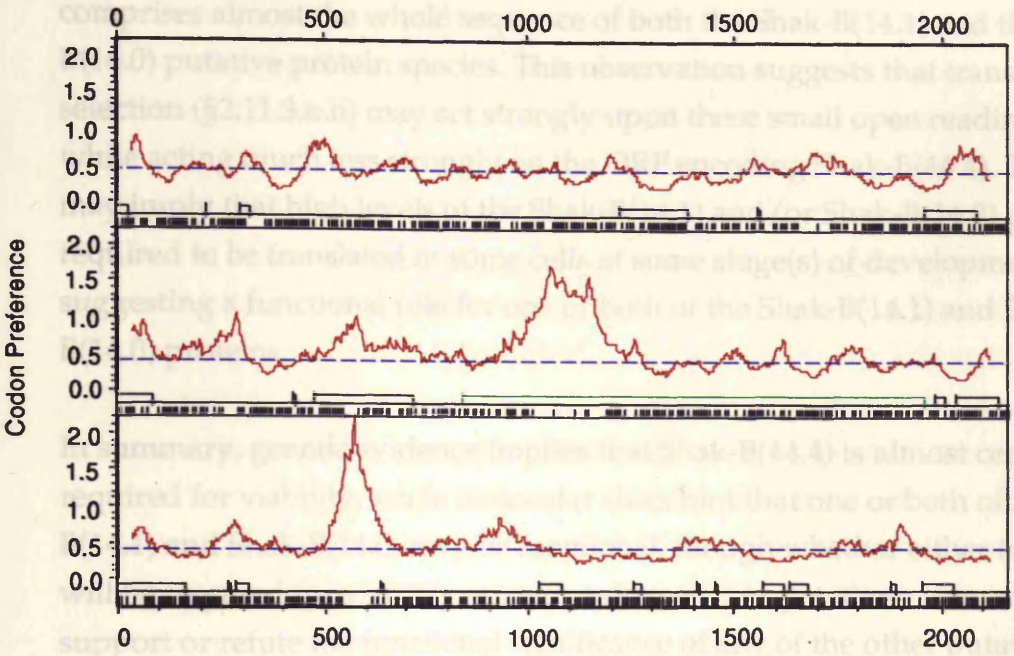


Figure 6.1: *Codonpreference* of the SIPC8 cDNA. The reading frame highlighted in green encodes Shak-B(44.4). It is apparent that towards the beginning of this reading frame the codon bias conforms well to that observed in highly biased *Drosophila* genes, while later this bias disappears. The reading frames of Shak-B(14.1) and Shak-B(14.0) are contained, almost entirely, within the highly biased region. See text, §2.11.3.a.ii and Figure 3.5 for further details.

It is clear from Figure 6.1 that codon usage towards the N terminus of the open reading frame of SIPC8 (encoding Shak-B(44.4)) conforms well to that observed in highly biased *Drosophila* genes (Shields, *et al.*, 1988). This N terminal high bias domain corresponds closely to the exon B derived coding region which comprises almost the whole sequence of both the Shak-B(14.1) and the Shak-B(14.0) putative protein species. This observation suggests that translational selection (§2.11.3.a.ii) may act strongly upon these small open reading frames, while acting much less strongly on the ORF encoding Shak-B(44.4). This in turn may imply that high levels of the Shak-B(14.1) and/or Shak-B(14.0) proteins are required to be translated in some cells at some stage(s) of development, suggesting a functional role for one or both of the Shak-B(14.1) and Shak-B(14.0) proteins.

In summary, genetic evidence implies that Shak-B(44.4) is almost certainly required for viability, while molecular clues hint that one or both of Shak-B(14.1) and Shak-B(14.0) may be functional, though whether either (or both) will be required for viability remains to be seen. As yet, there is no evidence to support or refute the functional significance of any of the other putative Shak-B(lethal) protein species described here.

6.2 CLUES TO THE FUNCTION(S) OF SHAKING-B PROTEINS

In chapter 3, I discussed the giant fibre phenotypes conferred by neural mutations at the *shaking-B* locus, and suggested that Shak-B(neural) protein(s) are required either for synaptic target selection by the TTMn and/or CGF neurons or for stabilisation of this synapse (§3.2.3). In this final section, I will integrate data already presented concerning the organisation of the *shaking-B* locus and the structures of its proposed protein products with functional clues from phenotypes conferred by mutations of *shak-B*, *l(1)ogre* and *unc-7* to further address the question of what the biochemical functions of Shak-B proteins and their relatives might be. While direct evidence of what Shak-B proteins do is a luxury which we lack, a thorough examination of indirect hints is valuable and, while we cannot expect concrete proofs from indirect data, we can at least arrive at testable hypotheses.

6.2.1 Clues from *unc-7*

Mutations at the *Caenorhabditis elegans unc-7* locus confer an uncoordinated phenotype. The bodies of the mutants show sharp, irregular bends, rather than the smooth, sinusoidal curves of wild type worms. This postural impairment ("kinking") is most prominent during forward motion, though it is also evident when the worms are moving backwards or are at rest. At the level of the light microscope, no morphological abnormalities which might explain this phenotype are apparent in muscle (Waterston, 1988) or neurons (W. Li and J. Shaw, unpublished; cited by Starich *et al.*, 1993). Under the electron microscope (J. White, N. Southgate and N. Thompson, unpublished; cited by Starich *et al.*, 1993) certain supernumerary gap junctions are detected. AVA interneurons normally form gap junctions with the DA, VA and AS motor neuron classes which drive backwards locomotion (Chalfie, *et al.*, 1985; Chalfie and White, 1988). Forward locomotion is normally driven by the DB and VB motor neurons which receive input, via gap junctions, from the AVB interneurons. In the *unc-7* mutants the AVA interneurons form gap junctions with the VB and DB motor neurons, in addition to serving their normal synaptic partners. Thus interneurons normally driving backwards movement are seen to connect with motor neurons causing forward movement, a miswiring which may be responsible for the kinking phenotype.

A fascinating parallel with the giant fibre mutations of *shaking-B* neural alleles is immediately apparent. *unc-7* mutant animals form inappropriate gap junctions between large interneurons and motor neurons, while *shaking-B*^{2} and *shaking-B*^{Passover} mutants fail to form gap junctions between large interneurons (the giant fibres) and motor neurons.

In order to elucidate the nature of the *unc-7* locus, a genetic and molecular analysis was carried out by Starich and coworkers (Starich, *et al.*, 1993). These experimenters isolated novel alleles of *unc-7* by transposon mutagenesis, and used these transposon tags to clone the locus. A 2.4kb cDNA containing all of the coding sequence of *unc-7* was isolated and sequenced and, as discussed above, its implied protein sequence is likely to have three or four transmembrane domains and is homologous to Shaking-B proteins and to Ogre.

In order to determine the tissue(s) in which *unc-7* function is required, Starich and colleagues performed an elegant mosaic analysis. The first mitosis in the developing *C. elegans* embryo is the division of the P0 cell into its daughters AB

and P1. The results of Starich *et al* demonstrated that when descendants of the P1 cell (including all body muscle cells but one) are hemizygous mutant for *unc-7*, phenotypically wild type worms result. Thus no requirement for *unc-7*$^+$ was found in muscle cells. All but two of the worm's nonpharyngeal neurons descend from the AB cell. In all mosaic worms showing the *unc-7* phenotype, the free duplication carrying *unc-7*$^+$ was seen to have been lost in the mitosis generating either the AB cell or its early descendants, leading to the conclusion that the requirement for *unc-7*$^+$ is diffusely spread among several descendants of AB. The AB cell divides into AB.a and AB.p. No mosaics were found in which duplication loss in the AB.a lineage alone was associated with an *unc-7* phenotype. Both AVA interneurons are descended from AB.al, a daughter of AB.a, thus a cell autonomous focus of the defect in the AVA interneurons was excluded. A caveat should be emphasised here: it was not at any point proven that the AVA miswiring phenotype is, in itself, sufficient to generate the *unc-7* uncoordination, and neither were worms with AB.a hemizygous mutant lineages studied by electron microscopy. It therefore remains a possibility that the miswiring of AVA interneurons is an AVA cell autonomous phenotype which alone is insufficient to generate uncoordination.

Starich and coworkers concluded that *unc-7* activity is required in several descendants of AB, perhaps uniquely in cells descended from AB.p. Having shown that *unc-7* activity is not required in muscle cells, these authors proposed, on the basis of the observed uncoordination, miswiring, and anaesthetic (see below) phenotypes that *unc-7* is required in neurons.

In addition to its role in coordinated locomotion, the *unc-7* product is also implicated in the worm's response to volatile anaesthetics (e.g. halothane and chloroform). Mutations in the genes *unc-79* and *unc-80* render worms hypersensitive to these anaesthetics (Sedensky and Meneely, 1987). Additional mutation of the *unc-7* locus suppresses this hypersensitivity (Morgan, *et al.*, 1990). The traditional view of the mechanism of action of volatile anaesthetics is that they exert an influence upon membrane ion channels indirectly, after binding to lipid components of membranes. Recent evidence challenges this view (Franks and Lieb, 1991; Matthews, 1992), supporting instead the contention that anaesthetic molecules act directly upon membrane ion channels. It has therefore been suggested that genes such as *unc-79*, *unc-80* and *unc-7* may encode ion channel subunits or proteins that interact with ion channels.

Further evidence that the product of *unc-7* might be involved in ion channel function comes from studies of the drug ivermectin. This is a semisynthetic derivative of the avermectins, a class of macrocyclic lactones originally isolated from *Streptomyces avermetilis*. The avermectins (and ivermectin) have been intensively studied, as they are potent antihelmintic and insecticidal drugs, yet have low toxicity to humans and mammals (Campbell, 1989). Thus they are widely used in vetinary medicine to control helminth infections and in humans to control *Onchocerca volvulus*, the causative agent of river blindness.

Arena and coworkers (Arena, *et al.*, 1991) used a *Xenopus* oocyte expression system to attempt to elucidate the target of the antihelmintic action of avermectin. *Xenopus laevis* oocytes were injected with 70 ng *C. elegans* mRNA and incubated for 2-3 days. Subsequent treatment with avermectin was found to increase the oocyte membrane permeability to chloride. This effect was not found in sham injected oocytes. These chloride currents could not be activated by changes in membrane potential, suggesting that avermectin activates a background and/or ligand gated chloride channel rather than a voltage gated one. Further investigations (Arena, *et al.*, 1992) demonstrated a glutamate sensitive current, which is almost certainly meditated by the same, avermectin sensitive, channel. The cloning, by expression in *Xenopus* oocytes, of two polypeptides meditating this glutamate/ivermectin sensitive chloride current (GluClα and GluClβ) has recently been reported in abstract form (Arena, *et al.*, 1994; Cully, *et al.*, 1994).

Intriguingly, mutations in the *unc-7* gene may render worms resistant to ivermectin (C. Johnson and P. Morgan, unpublished; cited by Starich *et al.*, 1993). Neither GluClα nor GluClβ shows any homology to Unc-7, however, being homologous instead to the glycine and GABA gated chloride ion channel subunits (reviewed by Betz, 1990). Thus the ivermectin-sensitive chloride channel can be reconstituted in oocytes without Unc-7 (GluClα and GluClβ alone being sufficient to form the ivermectin-sensitive channel), but the ivermectin resistance of *unc-7* mutants suggests that Unc-7 may, in some way, interact with this channel.

6.2.2 *l(1)ogre*

The first *l(1)ogre* mutation was identified as a mutation which impaired visual pattern recognition (Lipshitz and Kankel, 1985). One viable and three lethal *l(1)ogre* alleles have been recovered to date. Examination of both adults homozygous for the viable allele and rare adult escapers homozygous for lethal alleles reveal striking abnormalities in the optic lobes, but not elsewhere. The optic lobes of mutants are greatly reduced in size. Mosaic studies (Lipshitz and Kankel, 1985) show that defective vision, defective flight, and reduction in optic lobe size result if tissue in or near the optic lobe primordia is hemizygous mutant.

Light microscope examination of larvae reveals no abnormalities in mutants except for a reduction in the number of postembryonic neuroblasts and their offspring. The neuroblast population affected includes those of the optic formation centres which give rise to the adult optic lobes. Electron microscopy of mutants revealed extensive cell degeneration in the CNS of late larvae (Singh, *et al.*, 1989), in addition to more subtle structural abnormalities in muscles, and in imaginal discs and their derivatives. The principal *l(1)ogre* mutant phenotype was concluded to be a failure of generation and/or maintenance of postembryonic neuroblasts.

l(1)ogre was cloned by chromosome walking (Watanabe and Kankel, 1990). A 2.45 kb cDNA derived from the locus was cloned and sequenced, and its anticipated product was shown to be a 362 amino acid protein, which has since been found to be highly homologous with products of *shaking-B* (Crompton, *et al.*, 1992; this work) and is here proposed to have three or four transmembrane domains.

The pattern of expression of *l(1)ogre* transcripts has been studied in detail (Watanabe and Kankel, 1992). Transcripts are detected in the optic formation centres of larvae but are also observed in a large variety of other tissues and developmental stages. Watanabe and Kankel proposed that the discrepancy between the apparent requirement for Ogre function only in or around the optic formation centres and its widespread expression could be explained by the existence of a multigene family (also including *shaking-B*) rendering Ogre function redundant in many tissues where it is produced. Undoubtedly, Ogre and Shak-B proteins have high sequence homology and may have comparable biochemical functions but *shak-B* and *l(1)ogre* expression patterns show only limited overlap (e.g. *shak-B* expression is absent from the embryonic nervous

system- see below). Thus if the functional redundancy of *l(1)ogre* really is the explanation, then other members of this emerging protein family would have to be present to contribute to the apparent functional complementation occurring when Ogre function is absent.

Other explanations for the discrepancy between the observed expression patterns and the apparent functional requirement for Ogre protein are also possible, however. It has not been adequately proven that the strongest allele of *l(1)ogre* (*ogre*^{ljnl3}) is a null allele. Thus it may be that lower levels of Ogre function are required in tissues other than the optic lobe primordia and the levels provided by the existing hypomorphic alleles exceeds these lower thresholds, so not generating obvious phenotypes. A more extreme version of a similar argument states that the strongest *l(1)ogre* allele is, indeed, a null, but Ogre function is simply not required in some of the tissues where it is expressed (see Erickson, 1993, for a discussion of this phenomenon).

Expression was also studied at the protein level using polyclonal antisera raised against the C terminal EFAKQVEPSKHDRAK peptide. The homology with Shak-B proteins in this region of Ogre is minimal, and staining was shown to be attributable to the Ogre protein alone. Such studies revealed Ogre protein to be cytoplasmically distributed in imaginal disks, postembryonic neuroblasts and embryonic tissues, though whether the protein was located in the cytosol or bound to an organellar membrane could not be resolved at the resolution achieved.

Despite a substantial body of data concerning the nature of *l(1)ogre* mutant phenotypes, the structure of the *l(1)ogre* gene, and the expression of *l(1)ogre* transcripts and Ogre protein, the available data provide few good clues as to the possible biochemical function(s) of the Ogre protein.

6.2.3 Expression of *shaking-B*

The observed phenotypes of *shaking-B* mutants have been discussed earlier (§3.2.3; §3.3). An elegant series of *in situ* hybridisation studies has been performed by Martin Todman (Crompton, *et al.*, 1995), and is reviewed here.

6.2.3.a Embryonic *shak-B* expression

Expression of *shak-B* in embryos was initially studied using the KE2(1.8) cDNA (§3.5) as a probe. It subsequently became apparent that this cDNA overlaps with the P2.4 (*shak-B(neural)*) sequence by about 100 bases (in exon D, see Figure 4.8). The experiments were therefore repeated using the pLethal[1] (pL) probe (§2.3.3) which lacks this region of overlap. The results obtained using these different probes were indistinguishable, and reflect embryonic *shak-B(lethal)* expression. There is thought to be little or no embryonic expression of neural *shak-B* mRNA forms (Martin Todman, unpublished results). This finding is consistent with the absence of embryonic or larval phenotypes in *shak-B(neural)* mutants.

shak-B expression is not detectable in the embryonic nervous system, transcripts being found, instead, in derivatives of the mesoderm. A subset of the somatic muscle precursors, (and subsequently the muscles themselves) express *shak-B*, as do cells of the visceral mesoderm, the dorsal pharyngeal musculature, and the cardioblasts. Careful examination of these tissues in *shak-B(lethal)* embryos might be hoped to shed light on the nature of *shaking-B* lethality, and examination of the musculature and cardioblasts of these animals does indeed reveal minor defects (Emma Rushton and Mike Bate, personal communication). Thus, in both *shak-B*$^{R\text{-}9\text{-}29}$ and *shak-B*^{L41} homozygotes, some myoblasts occasionally fail to fuse with the developing muscle fibres (see Bate, 1990, for an account of muscle development during embryogenesis), while gaps are sometimes noted in the row of cardioblasts. These defects are, however, only minor, and no disruption whatsoever is detectable in gut morphology. It is therefore considered likely that lethality is caused by a functional defect in one or more of the tissues expressing *shak-B*, rather than being a consequence of the morphological disruptions seen.

6.2.3.b Expression of *shaking-B* transcripts in the pupal nervous system

Pupal expression was first examined using a probe, termed pNL, which detects both neural and lethal *shak-B* transcripts. At the end of the third larval instar, expression is seen in cells which will give rise to part of the suboesophageal ganglion of the adult and in a pair of cells in the central brain. At 12 hours after puparium formation (APF), this same pair of central brain cells is seen to stain much more strongly, and the position of these cells, just lateral to the midline,

---

1 This subclone was created, as part of this work, specifically for the purpose of monitoring expression of *shak-B(lethal)* transcripts.

at the posterior surface of the brain, suggests that they are the giant fibre somata. In addition, staining is prominent in a pair of unidentified cells in the suboesophageal ganglion, and in certain cells within the thoracic neuromeres of the ventral ganglion, including cells whose positions coincide with those of the PSI and TTMn cell bodies (§3.1).

At around 12 hr APF, faint *shak-B* expression begins to appear in most of the cells of the central brain and thoracic neuromeres of the ventral ganglion, this expression intensifying over the subsequent 15 hours of development. Between 25 and 30 hr APF, the optic lobes begin to express *shaking-B* and by 48 hr APF, most cells of the optic ganglia, brain and thoracic neuromeres express *shak-B* at high levels. This expression remains high for several hours before declining. By two days after eclosion, *shak-B* expression persists only in the putative CGF cell bodies.

Experiments with the pL probe do not reveal expression in the giant fibre cell bodies, or in the medulla or lamina of the optic lobes at any pupal or adult stages. The remaining cells which are found to hybridise with the pNL probe also stain with pL. These cells include the dorsal and lateral cells in the anterior mesothoracic neuromere believed to include the TTMn and/or PSI. Thus *shaking-B* expression in the giant fibres and in the lamina and medulla of the optic lobes is likely to be attributable to *shak-B(neural)* transcripts (as all *shak-B(lethal)* transcripts thus far identified (Figure 4.8) contain regions present in the pL probe), while the expression detected in other CNS cells, probably including cells postsynaptic to the giant fibres, is, at least in part, due to *shak-B(lethal)* transcripts.

Martin Todman's studies of *shak-B* transcript expression, as described in Crompton *et al.*, (1995) and summarised here, expands greatly upon a previous description of *shak-B* expression patterns (Krishnan, *et al.*, 1993). These latter authors described the expression of *shak-B* during the last quarter of pupal life (75hr APF onwards) and reported *shak-B* expression restricted to the CGF, TTMn and/or PSI, and a tiny handful of other cells close to the CGF cell bodies. Krishnan and colleagues were, at that time, only aware of the P2.4 transcript form, and did not declare which regions of this were used as probes in their experiments. If the probes used were derived only from regions upstream of the first common (*neural* and *lethal*) *shaking-B* exon (exon D), then this might explain the discrepancy between findings from the two studies. If this is,

indeed, the case, then *shak-B(neural)* expression in the pupal nervous system is rather limited when compared to that of *shak-B(lethal)*, suggesting a major role for *shak-B(lethal)* in the development of the imaginal nervous system. If homozygous *shak-B(lethal)* mutants can be rescued from late embryonic/early larval lethality using a SIPC8 transgene, then manipulating expression of the transgene during pupal development may allow further investigation of the nervous system role of *shak-B(lethal)*.

6.2.3.c Expression of Shaking-B proteins in the giant fibre system during metamorphosis

Marian Wilkin has raised Shak-B antisera against a synthetic peptide corresponding to 14 amino acids from the C-terminus of the Shak-B(44.4) and Shak-B(42.9) proteins (Phelan, *et al.*, 1996), and has used these to follow expression of Shaking-B proteins in the nervous system duing metamorphosis. Although the antisera detect both neural and lethal proteins, these may be distinguished by studying *shak-B*^{2} homozygotes, in which no Shak-B(42.9) protein is produced (§6.1.1). Expression is first detected in the nervous system at about 20 hours APF, at which stage both lethal and neural proteins are diffusely expressed at low levels, throughout almost all of the neuropil of the brain and thoracic ganglion. From about 30 to 40 hours APF, expression in four discrete regions of the mesothoracic neuromere of the thoracic ganglion becomes superimposed upon the low-level background expression. Two of these regions have been exactly identified by injecting lucifer yellow into the CGF prior to antibody staining. One is the terminal bend of the CGF, at precisely the point where it contacts the medial dendrite of the TTMn; another is the CGF-PSI synapse. The Shak-B expression pattern in the mesothoracic neuromere intensifies as development proceeds, and expression is maintained at high levels in the adult.

In the brain, only the low level neuropil expression is apparent until 50-60 hours APF, whereupon a broad cluster of punctate staining becomes apparent in the deutocerebrum. This is more intense in later pupae and adults. Lateral to this cluster, a number of processes are stained, which have been identified (using lucifer yellow filling of the CGF) as the regions where the dendrites of the CGFs and and CGIs arborise.

In *shak-B*^{2} mutants staining in the giant fibre system is lost, except at the region of the CGF-PSI synapse. Thus most of the giant fibre system Shak-B immunoreactivity is likely to be due to Shak-B(neural) while staining at the CGF-PSI synapse may be due, at least in part, to Shak-B(lethal) protein, an observation which is consistent with the transcript expression results described above. This observation, taken together with the fact that the PSI-CGF synapse is disrupted in *shak-B*^{2} mutants, suggests that neural and lethal protein forms colocalise at the CGF-PSI synapse.

Shak-B protein expression is also detected in other regions of the pupal nervous system, including the optic lobes. A full description of this expression is still to be completed (Marian B. Wilkin and Jane A. Davies, in preparation).

6.2.4 Shak-B, Ogre and Unc-7: A unifying theory?

Can we then assemble the substantial quantity of eclectic data about *shaking-B* and its homologues into a plausible unifying theory regarding the functions of this emerging gene family? I have suggested previously (§3.2.3), on the basis of the giant fibre mutant phenotypes conferred by *shak-B(neural)* alleles that Shak-B(neural) proteins might either be involved in target recognition (§1.2.2) by giant fibre system neurons or might encode proteins required for the formation of mature synapses in this system. These possibilities will be considered in turn.

6.2.4.a *shaking-B* is unlikely to encode target recognition molecules

The proposal that *shak-B(neural)* encodes a synaptic target recognition molecule was first made by Krishnan and colleagues (Krishnan, *et al.*, 1993). Having observed *shak-B* transcript expression in cells believed to be the giant fibre and TTMn, these authors speculated that Shak-B(42.9) might mediate synaptic target selection by a homophilic adhesion mechanism. However, while a number of molecules implicated in *Drosophila* nervous system development have been shown to mediate homophilic adhesion in tissue culture cells (e.g. Elkins, *et al.*, 1990; Kania, *et al.*, 1993; Nose, *et al.*, 1992; Snow, *et al.*, 1989), such proteins are all single transmembrane glycoproteins, whereas a three or four transmembrane domain structure is proposed here for Shak-B(42.9). The observed intracellular disribution of the Ogre protein is also hard to reconcile with any adhesive role.

In embryos, *shak-B(lethal)* is expressed in mesodermal derivatives, but *shak-B* is not expressed in embryonic neurons. Thus, if *shak-B* is active in target selection in embryonic neuromuscular development, then homophilic adhesion is not involved. Moreover, although low level Shak-B expression is seen in the giant fibre system at around the time the identified synapses are forming (Phelan, *et al.*, 1996), expression at this stage is low and only increases after synaptogenesis is underway. Thereafter expression is maintained at high levels into adulthood. This expression pattern is much more suggestive of a structural component of the synapse, rather than a target recognition molecule.

6.2.4.b Shak-B proteins are required for functional giant fibre system gap junctions

The data recently presented by Phelan and colleagues (Phelan, *et al.*, 1996) clearly demonstrate that Shak-B proteins are localised to gap junctions in the giant fibre system (discussed in §6.2.3.c). Moreover, these authors demonstrated by dye filling of the CGF, that the *shak-B*^{2} mutation completely abolishes dye coupling to other giant fibre system neurons, while having no detectable effect upon CGF morphology. It is therefore evident that Shak-B(42.9) is required for functional gap junctions between the CGF and the cells to which it is normally dye-coupled. Having argued, above, against a target selection role for Shak-B(42.9), two other major possibilities present themselves. Either Shak-B(42.9), (and presumably Shak-B(44.4) as well) might encode a gap junction channel molecule (as has been proposed by others: Barnes, 1994), or Shak-B proteins might play another role which is essential for the stable formation of gap junctions, and perhaps of other specialised membrane structures too. In order to address these two possibilities, I will discuss the following issues. Firstly, I will consider whether, in light of the structural analyses presented in chapter 5 and structural data concerning identified gap junction molecules, Shak-B proteins look like gap junction channel molecules. Secondly I will consider those aspects of expression patterns and mutant phenotypes of *shak-B* and its homologues which seem consistent with a gap junction hypothesis. Finally I will address those aspects of the data which seem inconsistent with a gap junction channel role. These argue in favour of the interpretation that Shak-B proteins in the giant fibre system act in some accessory role to support gap junctions and thus may act in other capacities at

other times in other tissues, perhaps in the organisation of other membrane specialisations.

6.2.4.c Might the Shak-B protein family be structural components of invertebrate gap junctions?

The molecules forming gap junctions in vertebrates, the Connexins, have been intensively studied (reviewed by Bennett, *et al.*, 1991). The Connexins are a superfamily of proteins with four transmembrane α-helices (Tibbitts, *et al.*, 1990), which, when expressed in *Xenopus* oocytes (Swenson, *et al.*, 1989) or in tissue culture cells (Eghbaldi, *et al.*, 1990) are sufficient to induce gap junction formation and intercellular coupling. Connexin molecules aggregate into hexamers, each hexamer forming a hemichannel. Hemichannels interact with hemichannels expressed by neighbouring cells to form full gap junctions spanning two bilayers. Hemichannels composed of one Connexin form may interact with hemichannels made of another Connexin, at least in some cases (e.g. Swenson, *et al.*, 1989).

For the sake of argument, let us temporarily indulge the speculation that Shak-B(44.4) and its homologues are *Drosophila* Connexins. There is no significant primary sequence homology between the Shaking-B protein family and the Connexins, but we might speculate instead that convergent evolution has created topologically homologous families of molecules lacking primary sequence homology but with similar biochemical functions.

Comparison of the structural features of the Shak-B and Connexin protein families reveals some striking similarities. Connexins, like the Shak-B family, are basic proteins. Connexins have four transmembrane domains (M1 to M4) with cytoplasmic N and C termini (Goodenough, *et al.*, 1988; Herzberg, *et al.*, 1988; Milks, *et al.*, 1988; Yancey, *et al.*, 1989), a topology identical to that proposed for the Shak-B family according to the 4TM model (§5.1.4.b). Within the Connexin family, sequence conservation is highest in the transmembrane regions and this is also true of the proposed TM domains of the Shak-B family (Figure 5.6), though the same observation applies also to several other ion channel families. In the Connexins, the extracellular domains show the next highest level of conservation (reflecting the fact that it is through these domains that Connexins of one hemichannel interact with those of another), while the

intracellular domains are poorly conserved. These comments are also true for the Shak-B family according to the 4TM model (Figure 5.6). The extracellular loops of the Connexins each have three perfectly conserved cysteine residues which are essential for Connexin function (Dahl, *et al.*, 1992). The proposed extracellular loops of the Shak-B family proteins, according to the 4TM model, each have two perfectly conserved cysteines (Figure 5.6). Connexins are non-glycosylated, and, according to the 4TM model, the Shak-B protein family would all lack functional glycosylation sites (table 5.A.3).

Of the four transmembrane helices present in the Connexins, the fourth is much more hydrophobic than the others while the third is strongly amphiphilic and has been proposed to be the channel-lining helix (Unwin, 1989). The third and fourth candidate TM helices in the Shak-B family share these features. In the M2 helix of the Connexins, a conserved proline residue is present (Suchyna, *et al.*, 1993). Proline is a helix-breaking residue, as rotation about its N-C_α bond is restricted, creating a fixed angle which forces a 15-20° bend in α helices (Barlow and Thornton, 1988; Deisenhofer, *et al.*, 1985) and the presence of proline in transmembrane helices has therefore been ascribed special significance (Brandl and Deber, 1986; von Heijne, 1991). In the case of the Connexins, the conserved proline in the M2 helix has been shown to play a crucial role in the voltage gating of the gap junction channel (Suchyna, *et al.*, 1993), while leaving other aspects of channel function unaltered. The proposed region B of the Shak-B family also has a conserved proline residue, though this is not proposed to be as central within the helix as that in M2 of the Connexins.

Despite these common features we might question the validity of ascribing a gap junctional role to the Shak-B family in the absence of primary sequence homology with the Connexins. With regard to this objection, it is noteworthy that, to date, no invertebrate molecule with such homology has been described, despite attempts to detect Connexin immunoreactivity and DNA sequences related to *connexin* genes in invertebrates (Berdan and Gilula, 1988; Ryerse, 1991). Furthermore, vertebrate cells in tissue culture will not form gap junctions with arthropod cells, despite the ability of cells from different vertebrate classes to establish functional gap junctions with each other (Epstein and Gilula, 1977). These lines of evidence hint that molecules with primary sequence homology to vertebrate Connexins might not exist in invertebrates. Consistent with this possibility are the observations that arthropod gap junctions have substantially different electrical (e.g. Verselis, *et al.*, 1991) and physical (e.g. Hanna, *et al.*,

1984) characteristics from their vertebrate counterparts. Given these differences, it is perhaps not unreasonable to propose that invertebrate gap junction proteins might not be closely related to Connexins in primary sequence.

If molecules resembling vertebrate Connexins are not present in invertebrates, then what are invertebrate gap junctions made of? To date, no sequences of *bona fide* invertebrate gap junction proteins have yet been presented. Attempts have been made to isolate such molecules from *Drosophila* (Ryerse, 1989), from the crayfish *Procambarus clarkii* (Berdan and Gilula, 1988), and from the lobster *Nephrops norvegicus* (Finbow, *et al.*, 1984), starting, in each case, with purification of proteins from gap junction-enriched membrane isolates. Efforts to isolate gap junction proteins from *Nephrops* resulted in the characterisation of a 16 kDa polypeptide, one member of an emerging channel protein family known as Ductins (Holzenburg, *et al.*, 1993). Ductins are intriguing four transmembrane helix proteins present in a wide variey of species and tissues, including *Nephrops* hepatopancreas, a variety of mouse tissues, chicken liver, and *Xenopus* liver (Finbow, *et al.*, 1993). In mouse, *Nephrops*, *Manduca sexta*, and *Drosophila*, the same protein mediates not only intercellular communication via gap-junction like structures ("pseudo gap junctions": Berdan and Gilula, 1988) but also functions as a subunit of the vacuolar H^+ATPase responsible for acidification of vacuolar cell compartments (Finbow, *et al.*, 1994).

The presumptive gap junction protein species isolated from *Drosophila* and from *Procambarus* are, unlike the Ductins, likely to be *bona fide* components of classical gap junctions, but to date these proteins have been little characterised. Five such protein species, with apparent molecular weights of 18, 26, 36, 52 and 54 kDa, were isolated from *Drosophila* larvae, and an antiserum raised against the 18kDa species was shown to label intercellular boundaries of wing imaginal discs and (at the electron microscope level) gap junctions in a gap junction-rich subcellular fraction.

None of the gap junction protein species identified thus far from *Drosophila* have sizes similar to those predicted for Ogre or Shak-B proteins, thus while Shak-B family proteins share some common structural themes with vertebrate gap junction molecules and it may be possible to speculate that these proteins form gap junction channels, it is not tenable to argue that these proteins alone are responsible for all of the fly's gap junctions.

6.2.4.d The gap junction channel hypothesis in light of expression patterns and mutant phenotypes

How does the proposal that Shak-B family proteins might encode structural components of gap junctions square with what we know of the expression patterns of these molecules and the mutant phenotypes resulting from their absence? One aspect of *shaking-B* expression which demands explanation is the widespread, transient and almost ubiquitous expression which occurs in the pupal nervous system midway through pupariation (Crompton, *et al.*, 1995); discussed in (§6.2.3.b). Is this consistent with widespread gap junction formation? While there are few data from *Drosophila* available to address this question, it is clear from vertebrate experiments that a transient phase of neuronal coupling is a very widespread feature in developing nervous systems (reviewed by Kandler and Katz, 1995), coupling being particularly widespread just before and during the initial period of neuronal circuit formation. It seems at least plausible that a similar high incidence of transient gap junction formation may occur in *Drosophila*.

Are the data concerning the *C. elegans unc-7* locus consistent with its product being a gap junction component? It is interesting that in vertebrates, gap junctional communication is known to be blocked at least in some tissues by volatile anaesthetics including halothane (Burt and Spray, 1989). If an equivalent effect occurred in *C. elegans*, we might anticipate that null *unc-7* mutants would show greater sensitivity to anaesthetic agents rather than a suppression of hypersensitivity. If Unc-7 genuinely is a gap junction component then the anaesthetic data would be more consistent with a mechanism whereby the primary target of the anaesthetic agent was a different ion channel and *unc-7* mutations, by reducing the electrical coupling between cells, served to insulate some cells from their intoxicated neighbours. Such a model is appealing because it predicts that mutations in *unc-7* would suppress the effects of anaesthetics upon multiple primary targets, and *unc-7* mutations are known to suppress the anaesthetic hypersensitivity phenotypes conferred by mutations in two distinct genes.

An analogous mechanism might be postulated to account for ivermectin resistance. As discussed earlier, ivermectin acts as an agonist of a glutamate-gated chloride channel. In many excitable cells the membrane potential is close to, or slightly less negative than, the equilibrium potential for chloride, hence

ivermectin would be anticipated to prevent membrane depolarisation and thus action potentials. Any cells electrically coupled to cells expressing the ivermectin receptor would be similarly silenced, thus a decline in gap junctional communication due to mutation of a gap junction channel gene might enhance resistance.

While these lines of evidence are broadly compatible with the hypothesis that the Shak-B protein family encodes gap junction channels, other data are not. Thus in the developing embryonic musculature of *Drosophila*, embryonic myotubes are electrically and dye coupled to their neighbours until between 13 and 13.25 hours after egg laying (AEL) whereupon they abruptly uncouple from each other (Broadie and Bate, 1993b; Gho, 1994). At this time, although muscle expression of *shak-B(lethal)* transcripts has peaked and is declining, expression is still detectable. This is inconsistent with *shak-B(lethal)* encoding the myotube coupling function unless translational regulation of *shak-B(lethal)* expression is also proposed. Furthermore, *shak-B(lethal)* is expressed in cardioblasts, which are not known to form gap junctions at any stage of their development (Tepass and Hartenstein, 1994). The existence of ectopic gap junctions in *unc-7* mutants is also hard to reconcile with a gap junction channel role for Unc-7, however, this same objection applies equally to any proposal of accessory involvement of Unc-7 in gap junction formation. Finally, the intracellular localisation of Ogre protein is not immediately consistent with a role in gap junction formation. On the other hand, in the case of Ductin, the same protein is expressed in gap-junction like plasma membrane structures and on cytoplasmic vacuolar membranes (Finbow, *et al.*, 1994), hence an intracellular localisation of Ogre does not completely refute the proposal that other members of the Shak-B family might encode gap junctional channels.

6.2.4.e On the functions of the Shaking-B family: A summary

In summary, the structural features of the Shaking-B protein family are, according to the proposals made in Chapter 5, somewhat reminiscent of those of vertebrate Connexins, suggesting that the Shaking-B family might encode gap junction channel components. While some of the phenotypic and expression data regarding *shaking-B* and its homologues are broadly consistent with such an interpretation, there is also a substantial weight of data to confound this hypothesis. If the Shak-B family do not encode gap junction channels, they may play an accessory role in their organisation, and in that of

other membrane specialisations; indeed other molecules believed to have four transmembrane domains are known to play roles in the organisation of specialised membrane structures, e.g. tight junctions (Furuse, *et al.*, 1993) and urothelium (Yu, *et al.*, 1994).

There remains a huge number of unresolved questions regarding *shaking-B* and its homologues. The structural models proposed in Chapter 5 demand rigorous experimental testing, as does the proposal that Shaking-B proteins might encode gap junction channels. If the gap junction channel hypothesis is incorrect then the urgent question of what Shaking-B proteins and their homologues actually do remains open. It is my hope that the molecular analysis presented here will provide a solid and informative foundation for future experiments to address these and other questions about the functions of *shaking-B*, and its role in nervous system development.

References

Alberts, B., D. Bray, J. Lewis, M. Raff, K. Roberts and J. D. Watson, (1989). Molecular Biology of the Cell. (New York and London : Garland).

Anderson, M. S., M. E. Halpern and H. Keshishian, (1988). Identification of the neuropeptide transmitter proctolin in *Drosophila* larvae. Characterisation of muscle fiber-specific neuromuscular endings. J. Neurosci. **8**, 242-255.

Arena, J. P., K. K. Liu, P. P. Paress and D. F. Cully, (1991). Avermectin-sensitive chloride currents induced by *Caenorhabditis elegans* RNA in *Xenopus* oocytes. Mol. Pharmacol. **40**, 368-374.

Arena, J. P., K. K. Liu, P. S. Paress, J. M. Schaeffer and D. F. Cully, (1992). Expression of a glutamate-activated chloride current in *Xenopus* oocytes injected with *Caenorhabditis elegans* RNA: evidence for modulation by ivermectin. Mol. Brain Res. **15**, 339-348.

Arena, J. P., K. K. Liu, D. K. Vassilatis, P. S. Paress and D. F. Cully, (1994). Properties of a novel glutamate-gated channel from *Caenorhabditis elegans*. Biophys. J. **66**, 380.

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

Bacon, J. P. and N. J. Strausfeld, (1986). The dipteran 'Giant fibre' pathway: neurons and signals. J. Comp. Physiol. **158**, 529-548.

Baird, D. H., M. Koto and R. J. Wyman, (1993). Dendritic reduction in *Passover*, a *Drosophila* mutant with a defective giant fibre neuronal pathway. J. Neurobiol. **24**, 971-984.

Baird, D. H., A. P. Schalet and R. J. Wyman, (1990). The *Passover* locus in *Drosophila melanogaster*: Complex complementation and different effects on the giant fibre neural pathway. Genetics **126**, 1045-1059.

Bairoch, A., (1991). PROSITE: a dictionary of sites and patterns in proteins. Nucl. Acids Res. **19 (Supplement)**, 2241-2245.

Balakrishnan, R. and V. Rodrigues, (1991). The *Shaker* and *shaking-B* genes specify elements in the processing of gustatory information in *Drosophila melanogaster*. J. Exp. Biol. **157**, 161-181.

Bangham, J. A., (1988). Data-sieving hydrophobicity plots. Anal. Biochem. **174**, 142-145.

Barlow, D. J. and J. M. Thornton, (1988). Helix geometry in proteins. J. Mol. Biol. **201**, 601-619.

Barnes, T. M., (1994). OPUS: A growing family of gap junction proteins? TIG **10**, 303-305.

Barrantes, F. J., (1975). The nicotinic cholinergic receptor: different compositions evidenced by statistical analysis. Biochim. Biophys. Res. Commun. **62**, 407-414.

Bastiani, M. J. and C. S. Goodman, (1986). Guidance of neuronal growth cones in the grasshopper embryo. III. Recognition of specific glial pathways. J. Neurosci. **6**, 3542-3551.

Bastiani, M. J., A. L. Harrelson, P. M. Snow and C. S. Goodman, (1987). Expression of Fasciclin I and II glycoproteins on subsets of axon pathways during neuronal development in the grasshopper. Cell **48**, 745-755.

Bastiani, M. J., J. A. Raper and C. S. Goodman, (1984). Pathfinding by neuronal growth cones in grasshopper embryos. III. Selective affinity of the G growth cone for the P cells within the A/P fascicle. J. Neurosci. **4**, 2311-2328.

Bate, C. M., (1976). Pioneer neurones in an insect embryo. Nature **260**, 54-56.

Bate, C. M., (1990). The embryonic development of larval muscles in *Drosophila*. Development **110**, 791-804.

Bate, C. M., E. Rushton and D. Currie, (1991). Cells with persistent *twist* expression are the embryonic precursors of adult muscles in *Drosophila*. Development **113**, 79-89.

Bell, L. R., E. M. Maine, P. Schedl and T. W. Cline, (1988). *Sex-lethal*, a *Drosophila* sex determination switch gene, exhibits sex-specific RNA splicing and sequence similarity to RNA binding proteins. Cell **55**, 1037-1046.

Bellen, H. J., S. Kooyer, D. D'Evelyn and J. Perlman, (1992). The *Drosophila Couch potato* protein is expressed in nuclei of peripheral neuronal precursors and shows homology to RNA-binding proteins. Genes Dev. **6**, 2125-2136.

Bender, W., P. Spierer and D. S. Hogness, (1983). Chromosomal walking and jumping to isolate DNA from the *Ace* and *rosy* loci and the *Bithorax* complex in *Drosophila melanogaster*. J. Mol. Biol. **168**, 17-33.

Bennett, M. V. L., L. C. Barrio, T. A. Bargiello, D. C. Spray, E. Hertzberg and J. C. Saez, (1991). Gap junctions: New tools, new answers, new questions. Neuron **6**, 305-320.

Bentley, D. and H. Keshishian, (1982). Pioneer neurons and pathways in insect appendages. TINS **5**, 354-358.

Bentley, D. and H. Keshishian, (1983). Pioneer axons lose directed growth after selected killing of guidepost cells. Nature **304**, 62-65.

Berdan, R. C. and N. B. Gilula, (1988). The arthropod gap junction and pseudo-gap junction: Isolation and preliminary biochemical analysis. Cell Tissue Res. **251**, 257-274.

Betz, H., (1990). Ligand-gated ion channels in the brain: the amino acid receptor superfamily. Neuron **5**, 383-392.

Bieber, A. J., P. M. Snow, M. Hortsch, N. H. Patel, J. R. Jacobs, Z. R. Traquina, J. Schilling and C. S. Goodman, (1989). *Drosophila neuroglian*: A member of the immunoglobulin superfamily with extensive

homology to the vertebrate neural adhesion molecule L1. Cell **59**, 447-460.

Biou, V., J.-F. Gibrat, J. M. Levin, B. Robson and J. Garnier, (1988). Secondary structure prediction: combination of three different methods. Protein Engineering **2**, 185-191.

Birnboim, H. C. and J. Doly, (1979). A rapid alkaline extraction procedure for screening recombinant plasmid DNA. Nucl. Acids Res. **7**, 1513-1523.

Bodmer, M. and M. Ashburner, (1984). Conservation and change in DNA sequences coding for alcohol dehydrogenase in sibling species of *Drosophila*. Nature **309**, 425-430.

Brandl, C. J. and C. M. Deber, (1986). Hypothesis about the function of membrane-buried proline residues in transport proteins. Proc. Natl Acad. Sci. U. S. A. **83**, 917-921.

Breslauer, K. J., R. Frank, H. Blocker and L. A. Markey, (1986). Predicting DNA duplex stability from the base sequence. Proc. Natl Acad. Sci. U. S. A. **83**, 3746-3750.

Broadie, K. and C. M. Bate, (1993a). Activity-dependent development of the neuromuscular synapse during *Drosophila* embryogenesis. Neuron **11**, 607-619.

Broadie, K. and C. M. Bate, (1993b). Development of larval muscle properties in the embryonic myotubes of *Drosophila melanogaster*. J. Neurosci. **13**, 167-180.

Broadie, K. and C. M. Bate, (1993c). Development of the embryonic neuromuscular synapse of *Drosophila melanogaster*. J. Neurosci. **13**, 144-166.

Broadie, K. and C. M. Bate, (1993d). Innervation directs receptor synthesis and localisation in *Drosophila* embryo synaptogenesis. Nature **361**, 350-352.

Broadie, K., H. Sink, D. VanVactor, D. Fambrough, P. M. Whitington, C. M. Bate and C. S. Goodman, (1993). From growth cone to synapse: the life history of the RP3 motor neuron. Development **1993 Supplement**, 227-238.

Broadie, K. S., (1994). Synaptogenesis in *Drosophila*- Coupling genetics and electrophysiology. J. Physiol. (Paris) **88**, 123-139.

Brookfield, J., (1992). Can genes be truly redundant? Curr. Biol. **2**, 553-554.

Brow, M. A. D., (1990). Sequencing with Taq DNA polymerase. In PCR Protocols: a guide to methods and applications, M. A. Innis, D. H. Gelfand, J. J. Sninsky and T. J. White, eds. (San Diego: Academic Press), pp. 189-196.

Brown, N. H. and F. C. Kafatos, (1988). Functional cDNA libraries from *Drosophila* embryos. J. Mol. Biol. **203**, 425-437.

Budnik, V., Y. Zhong and C. -F. Wu, (1990). Morphological plasticity of motor axons in *Drosophila* mutants with altered excitability. J. Neurosci. **10**, 3754-3768.

Bullock, W. O., J. M. Fernandez and J. M. Short, (1987). XL1-Blue: A high efficiency plasmid transforming *recA Escherichia coli* strain with β-galactosidase selection. Biotechniques **5**, 376-379.

Burns, F. R., S. v. Kannen, L. Guy, J. A. Raper, J. Kamholz and S. Chang, (1991). DM-GRASP, a novel immunoglobulin superfamily axonal surface protein that supports neurite extension. Neuron **7**, 209-220.

Burt, J. M. and D. C. Spray, (1989). Volatile anesthetics block intercellular communication between neonatal rat myocardial cells. Circulation Research **65**, 829-837.

Campbell, W. C., (1989). Ivermectin and abamectin. (New York : Springer-Verlag).

Caruthers, M. H., (1985). Gene synthesis machines: DNA chemistry and its uses. Science **230**, 281-285.

Cavener, D. and S. C. Ray, (1991). Eukaryotic start and stop translation sites. Nucl. Acids Res. **19**, 3185-3192.

Cavener, D. R., (1987). Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. Nucl. Acids Res. **15**, 1353-1361.

Chalfie, M., J. E. Sulston, J. G. White, E. Southgate, J. N. Thompson and S. Brenner, (1985). The neural circuit for touch sensitivity in *Caenorhabditis elegans*. J. Neurosci. **5**, 956-964.

Chalfie, M. and J. White, (1988). The nervous system. In The nematode *Caenorhabditis elegans*, W. B. Wood, ed. (Cold Spring Harbor, N. Y.: Cold Spring Harbor Laboratory Press).

Chiba, A., P. Snow, H. Keshishian and Y. Hotta, (1995). Fasciclin III as a synaptic target recognition molecule in *Drosophila*. Nature **374**, 166-168.

Chothia, C., (1976). The nature of accessible and buried surfaces in proteins. J. Mol. Biol. **105**, 1-14.

Chou, P. Y. and G. D. Fasman, (1978). Prediction of the secondary structure of proteins from their amino acid sequence. Advan. Enzymol. **47**, 45-148.

Colamarino, S. A. and M. Tessier-Lavigne, (1995). The axonal chemoattractant Netrin-1 is also a chemorepellent for trochlear motor axons. Cell **81**, 621-629.

Covarribius, L. and F. Bolivar, (1982). Construction and characterisation of new cloning vehicles VI: Plasmid pBR329, a new derivative of pBR328 lacking the 428 base pair inverted duplication. Gene **17**, 79-89.

Crompton, D. E., A. Griffin, J. A. Davies and G. L. G. Miklos, (1992). Analysis of a cDNA from the neurologically active locus *shaking-B* (*Passover*) of *Drosophila melanogaster*. Gene **122**, 385-386.

Crompton, D. E., M. Todman, M. Wilkin, S. Ji and J. Davies, (1995). Essential and neural transcripts from the *Drosophila shaking-B* locus are differentially expressed in the embryonic mesoderm and pupal nervous system. Dev. Biol. **170**, 142-158.

Cully, D. F., D. K. Vassilatis, P. S. Paress, K. K. Liu and J. P. Arena, (1994). Expression cloning of a novel glutamate-gated chloride channel from *Caenorhabditis elegans*. Biophys. J. **66**, 380.

Dahl, G., R. Werner, E. Levine and C. Rabadan-Diehl, (1992). Mutational analysis of gap junction formation. Biophys. J. **62**, 172-180.

Deisenhofer, J., O. Epp, K. Miki, R. Huber and H. Michel, (1985). Structure of the protein subunits in the photosynthetic reaction centre of *Rhodopseudomonas viridis* at 3Å resolution. Nature **318**, 618-624.

Devereux J., P. Haeberli and O. Smithies, (1984). A comprehensive set of sequence analysis programs for the VAX. Nucl. Acids Res. **12**, 387-395.

DeWitt, D. L. and W. L. Smith, (1988). Primary structure of prostaglandin G/H synthase from sheep vesicular gland determined from the complementary DNA sequence. Proc. Natl Acad. Sci. U. S. A. **85**, 1412-1416.

Dodd, J. and T. M. Jessell, (1988). Axon guidance and the patterning of neuronal projections in vertebrates. Science **242**, 692-699.

Dodd, J., S. B. Morton, D. Karagogeous, M. Yamamoto and T. M. Jessell, (1988). Spatial regulation of axonal glycoprotein expression on subsets of embryonic spinal neurons. Neuron **1**, 105-116.

Doe, C. Q., Q. ChuLa-Graff, D. M. Wright and M. P. Scott, (1991). The *prospero* gene specifies cell fates in the *Drosophila* central nervous system. Cell **65**, 451-464.

Doolittle, R., (1981). Similar amino acid sequences: Chance or common ancestry? Science **214**, 149.

Dover, G. A., (1993). Evolution of genetic redundancy for advanced players. Curr. Opin. Genet. Dev. **3**, 902-910.

du Lac, S., M. J. Bastiani and C. S. Goodman, (1986). Guidance of neuronal growth cones in the grasshopper embryo. II. Recognition of a specific axonal pathway by the aCC neuron. J. Neurosci **6**, 3532-3541.

Eghbaldi, B., J. A. Kessler and D. C. Spray, (1990). Expression of gap junction channels in communication-incompetent cells after stable transfection with a cDNA encoding Connexin 32. Proc. Natl Acad. Sci. U. S. A. **87**, 1328-1331.

Eisenberg, D., R. M. Weiss and T. C. Terwilliger, (1982a). The helical hydrophobic moment: a measure of the amphiphilicity of a helix. Nature **299**, 371-374.

Eisenberg, D., R. M. Weiss, T. C. Terwilliger and W. Wilcox, (1982b). Faraday Symp. Chem. Soc. **17**, 109-120.

Eisenberg, D., E. Schwarz, M. Komaromy and R. Wall, (1984). Analysis of membrane and surface protein sequences with the hydrophobic moment plot. J. Mol. Biol. **179**, 125-142.

Elkins, T., M. Hortsch, A. J. Bieber, P. M. Snow and C. S. Goodman, (1990). *Drosophila fasciclin I* is a novel homophilic adhesion molecule that along with *fasciclin III* can mediate cell sorting. J. Cell Biol. **110**, 1825-1832.

Elkins, T., K. Zinn, L. McAllister, F. M. Hoffman and C. S. Goodman, (1990). Genetic analysis of a *Drosophila* cell-adhesion molecule: Interaction of *fasciclin I* and *Abelson* tyrosine kinase mutations. Cell **60**, 565-575.

Engelman, D. M. and T. A. Steitz, (1981). The spontaneous insertion of proteins into and across membranes: The helical hairpin hypothesis. Cell **23**, 411-422.

Engelman, D. M., T. A. Steitz and A. Goldman, (1986). Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Ann. Rev. Biophys. Biophys. Chem. **15**, 321-353.

Epstein, M. L. and N. B. Gilula, (1977). A study of communication specificity between cells in culture. J. Cell Biol. **75**, 769-787.

Erickson, H. P., (1993). Gene knockouts of *c-src*, *transforming growth factor β1*, and *tenascin* suggest superfluous, non-functional expression of proteins. J. Cell Biol. **120**, 1079-1081.

Faissner, A. and J. Kruse, (1990). J1/Tenascin is a repulsive substrate for central nervous system neurons. Neuron **5**, 627-637.

Fan, J. and J. A. Raper, (1995). Localized collapsing cues can steer growth cones without inducing their full collapse. Neuron **14**, 263-274.

Fasman, G. and W. A. Gilbert, (1990). The prediction of transmembrane protein sequences and their conformation: an evaluation. TIBS **15**, 89-92.

Feinberg, A. P. and B. Vogelstein, (1983). A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. **132**, 6-13.

Feinberg, A. P. and B. Vogelstein, (1984). Addendum: A technique for radiolabelling DNA restriction endonuclease fragments to high specific activity. Anal. Biochem. **137**, 266-267.

Feng, Y., L. E. Gunter, E. L. Organ and D. R. Cavener, (1991). Translation initiation in *Drosophila melanogaster* is reduced by mutations upstream of the AUG initiator codon. Mol. Cell. Biol. **11**, 2149-2153.

Fickett, J. W., (1982). Recognition of protein coding regions in DNA sequences. Nucl. Acids Res. **10**, 5303-5318.

Finbow, M. E., T. Eldridge, J. Bumltjens, N. J. Lane, J. Shuttleworth and J. D. Pitts, (1984). Isolation and characterisation of arthropod gap junctions. EMBO J. **3**, 2271-2278.

Finbow, M. E., S. F. Goodwin, L. Meagher, N. J. Lane, J. Keen, J. B. C. Findlay and K. Kaiser, (1994). Evidence that the 16 kDa proteolipid (subunit c) of the vacuolar H^+-ATPase and Ductin from gap junctions are the same polypeptide in *Drosophila* and *Manduca*: Molecular cloning of the *Vha16k* gene from *Drosophila*. J. Cell Sci. **107**, 1817-1824.

Finbow, M. E., S. John, E. Kam, D. K. Apps and J. D. Pitts, (1993). Disposition and oritentation of Ductin (DCCD-reactive vacuolar H^+ ATPase subunit) in mammalian membrane complexes. Exp. Cell Res. **207**, 261-270.

Finer-Moore, J. and R. M. Stroud, (1984). Amphipathic analysis and possible formation of the ion channel in an acetylcholine receptor. Proc. Natl Acad. Sci. U. S. A. **81**, 155-159.

Fitzgerald, M., G. C. Kwiat, J. Middleton and A. Pini, (1993). Ventral spinal cord inhibition of neurite outgrowth from embryonic rat dorsal root ganglia. Development **117**, 1377-1384.

Franks, N. P. and W. R. Lieb, (1991). Stereospecific effects of inhalational general anaesthetic optical isomers on nerve ion channels. Science **254**, 427-430.

Freier, S. M., R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson and D. H. Turner, (1986). Improved free energy parameters for predictions of RNA duplex stability. Proc. Natl. Acad. Sci. U. S. A. **83**, 9373-9377.

Fuchs, R., (1991). MacPattern: protein pattern searching on the Apple Macintosh. CABIOS **7**, 105-106.

Fuchs, R., P. Stoer, P. Rice, R. Omond and G. Cameron, (1990). New services of the EMBL data library. Nucl. Acids Res. **18**, 4319-4323.

Furuse, M., T. Hirase, M. Itoh, A. Nagafuchi, S. Yonemura, S. Tsukita and S. Tsukita, (1993). Occludin: A novel integral membrane protein localising at tight junctions. J. Cell Biol. **123**, 1777-1788.

Gait, M. J., (1990). Chemical Synthesis. In Nucleic Acids in Chemistry and Biology, G. M. Blackburn and M. J. Gait, eds. (Oxford: IRL Press).

Gertler, F. B., R. L. Bennett, M. J. Clark and F. M. Hoffmann, (1989). *Drosophila abl* tyrosine kinase in embryonic CNS axons: a role in axonogenesis is revealed through dosage-sensitive interactions with *disabled*. Cell **58**, 103-113.

Gho, M., (1994). Voltage-clamp analysis of gap junctions between embryonic muscles in *Drosophila*. J. Physiol. (London) **481**, 371-383.

Ghysen, A., C. Dambly-Chaudiere, E. Aceves, L. Jan and Y. Jan, (1986). Sensory neurons and peripheral pathways in *Drosophila* embryos. Roux's Arch. Dev. Biol. **195**, 281-289.

Ghysen, A. and R. Janson, (1980). Sensory pathways in *Drosophila*. In Development and Neurobiology of *Drosophila*, O. Siddiqi, P. Babu, L. Hall and J. Hall, eds. (New York: Plenum).

Gierasch, L. M., (1989). Signal Sequences. Biochemistry **28**, 923-930.

Glass, D. B., M. R. El-Maghrabi and S. J. Pilkis, (1986). Synthetic peptides corresponding to the site phosphorylated in 6-phosphofructo-2-kinase/fructose-2,6-bisphosphatase as substrates of cyclic nucleotide-dependent protein kinases. J. Biol. Chem. **261**, 2987-2993.

Gloor, G. B., N. A. Nassif, D. M. Johnson-Schlitz, C. R. Preston and W. R. Engels, (1991). Targeted gene replacement in *Drosophila* via P element induced gap repair. Science **253**, 1110-1117.

Goodenough, D. A., D. L. Paul and L. Jesaitis, (1988). Topological distribution of two connexin32 antigenic sites in intact and split rodent hepatocyte gap junctions. J. Cell Biol. **107**, 1817-1824.

Goodman, C. S., M. J. Bastiani, C. Q. Doe, S. du Lac, S. L. Helfand, J. Y. Kuwada and J. B. Thomas, (1984). Cell recognition during neuronal development. Science **225**, 1271-1279.

Goodman, C. S. and C. J. Shatz, (1993). Developmental mechanisms that generate precise patterns of neuronal connectivity. Cell/Neuron Supplement **Cell Vol. 72/Neuron Vol. 10**, 77-98.

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

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

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

Gray, T. M. and B. W. Matthews, (1984). Intrahelical hydrogen bonding of serine, threonine and cysteine residues within α helices and its relevance to membrane-bound proteins. J. Mol. Biol. **175**, 75-81.

Greeningloh, G., A. J. Bieber, E. J. Rehm, P. M. Snow, Z. R. Traquina, M. Hortsch, N. H. Patel and C. S. Goodman, (1990). Molecular genetics of neuronal recognition in *Drosophila*: evolution and function of immunoglobulin superfamily cell adhesion molecules. C. S. H. S. Q. B. **LV**, 327-340.

Gribskov, M., J. Devereux and R. R. Burgess, (1984). The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression. Nucl. Acids Res. **12**, 539-549.

Grunstein, M. and D. S. Hogness, (1975). Colony hydridisation: A method for the isolation of cloned DNAs that contain a specific gene. Proc. Natl Acad. Sci. U. S. A. **72**, 3961.

Guthrie, S. and A. Pini, (1995). Chemorepulsion of developing motor axons by the floor plate. Neuron **14**, 1117-1130.

Gyllensten, U. B. and H. A. Ehrlich, (1988). Generation of single stranded DNA by the polymerase chain reaction and its application to the direct sequencing of the *HLA-DQA* locus. Proc. Natl. Acad. Sci. U. S. A. **85**, 7652-7656.

Halpern, M. E., M. S. Anderson, J. Johansen and H. Keshishian, (1988). Octopamine immunoreactive nerve terminals are found on a single identified muscle fiber of the *Drosophila* larval body wall. Soc. Neurosci. Abstr. **14**, 383.

Halpern, M. E., A. Chiba, J. Johansen and H. Keshishian, (1991). Growth cone behaviour underlying the development of stereotypic synaptic connections in *Drosophila* embryos. J. Neurosci. **11**, 3227-3238.

Hanna, R. B., G. D. Pappas and M. V. L. Bennett, (1984). The fine structure of identified electrotonic synapses following increased coupling resistance. Cell Tissue Res. **235**, 243-249.

Harrelson, A. L., (1992). Molecular mechanisms of axon guidance in the developing insect nervous system. J. Exp. Zool. **261**, 310-321.

Harrelson, A. L. and C. S. Goodman, (1988). Growth cone guidance in insects: Fasciclin II is a member of the immunoglobulin superfamily. Science **242**, 700-708.

Hartmann, E., T. A. Rapoport and H. F. Lodish, (1989). Predicting the orientation of eukaryotic membrane-spanning proteins. Proc. Natl Acad. Sci. U. S. A. **86**, 5786-5790.

Hedgecock, E. M., J. G. Culotti and D. H. Hall, (1990). The *unc-5*, *unc-6*, and *unc-40* genes guide circumferential migrations of pioneer axons and mesodermal cells on the epidermis in *C. elegans*. Neuron **2**, 61-85.

Heery, D. M., F. Gannon and R. Powell, (1990). A simple method for subcloning DNA fragments from gel slices. TIG **6**, 173.

Heffner, C. D., A. G. S. Lumsden and D. D. M. O'Leary, (1990). Target control of collateral extension and directional axon growth in the mammalian brain. Science **247**, 217-220.

Hegner, M., A. v. Kieckebusch-Gück, R. Falchetto, P. James, G. Semenza and N. Mantei, (1992). Single amino acid substitutions can convert the uncleaved signal-anchor of sucrase-isomaltase to a cleaved signal sequence. J. Biol. Chem. **267**, 16928-16933.

Heisenberg, M., (1971). Separation of receptor and lamina potentials in the electroretinogram of normal and mutant *Drosophila*. J. Exp. Biol. **55**, 85-100.

Henderson, R., J. M. Baldwin and T. A. Ceska, (1990). Model for the structure of bacteriorhodopsin based on high resolution electron cryo-microscopy. J. Mol. Biol. **213**, 899-929.

Henikoff, S., M. A. Keene, K. Fechtel and J. W. Fristrom, (1986). Gene within a gene: Nested *Drosophila* genes encode unrelated proteins on opposite DNA strands. Cell **44**, 33-42.

Herzberg, E. L., R. M. Disher, A. A. Tiller, Y. Zhou and R. G. Cook, (1988). Topology of the Mr 27000 liver gap junction protein. Cytoplasmic localization of amino and carboxy termini and a hydrophilic domain which is protease sensitive. J. Biol. Chem. **263**, 19105-19111.

Hogue, B. G. and D. P. Nayak, (1994). Deletion mutation in the signal anchor domain activates cleavage of the influenza virus neuraminidase, a type II transmembrane protein. J. Gen. Virol. **75**, 1015-1022.

Holzenburg, A., P. C. Jones, T. Franklin, T. Pali, T. Heimburg, D. Marsh, J. B. C. Findlay and M. E. Finbow, (1993). Evidence for a common structure for a class of membrane channels. Eur. J. Biochem. **213**, 21-30.

Homyk, T., J. Szidonya and D. T. Suzuki, (1980). Behavioral mutants of *Drosophila melanogaster* III. Isolation and mapping of mutations by direct visual observation of behavioral phenotypes. Mol. Gen. Genet. **177**, 553-565.

Hong, W. and D. Doyle, (1990). Molecular dissection of the NH_2-terminal signal/anchor sequence of rat dipeptidyl peptidase IV. J. Cell. Biol. **111**, 323-328.

Honig, B. H. and W. L. Hubbell, (1984). Stability of "salt bridges" in membrane proteins. Proc. Natl Acad. Sci. U. S. A. **81**, 5412-5416.

Honig, B. H., W. L. Hubbell and R. F. Flewelling, (1986). Electrostatic interactions in membranes and proteins. Ann. Rev. Biophys. Biophys. Chem. **15**, 163-193.

Hunter, T., (1982). Synthetic peptide substrates for a tyrosine protein kinase. J. Biol. Chem. **257**, 4843-4848.

Hutter, P. and M. Ashburner, (1987). Genetic rescue of inviable hybrids between *Drosophila melanogaster* and its sibling species. Nature **327**, 331-333.

Hutter, P., J. Roote and M. Ashburner, (1990). A genetic basis for the inviability of hybrids between sibling species of *Drosophila*. Genetics **124**, 909-920.

Ikeda, K. and J. H. Koenig, (1988). Morphological identification of the motor neurons innervating the dorsal longitudinal flight muscles of *Drosophila melanogaster*. J. Comp. Neurol. **273**, 436-444.

Ikeda, K., J. H. Koenig and T. Tsuruhara, (1980). Organisation of identified axons innervating the dorsal longitudinal flight muscles of *D. melanogaster*. J. Neurocytol. **9**, 799-823.

Ikemura, T., (1981). Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: A proposal for a synonymous codon choice that is optimal for the *E. coli* translation system. J. Mol. Biol. **151**, 389-409.

Innis, M. A., D. H. Gelfand, J. J. Sninsky and T. J. White, (1990). PCR Protocols: A Guide to Methods and Applications. (San Diego : Academic Press Inc.).

Jähnig, F., 1990. Structure predictions of membrane proteins are not that bad. TIBS **15**, 93-95.

Jarecki, J. and H. Keshishian, (1993). Blocking synaptic activity during neuromuscular synaptogenesis promotes collateral sprouting in *Drosophila* embryos. Soc. Neurosci. Abstr. **19**, 645.

Jennings, M., (1989). Topography of membrane proteins. Ann. Rev. Biochem. **58**, 999-1027.

Jessell, T. M. and E. R. Kandel, (1993). Synaptic transmission: A bidirectional and self-modifiable form of cell-cell communication. **Cell, Vol 72/ Neuron, Vol. 10 (Suppl.)** 1-30.

Jia, X., M. Gorczyca and V. Budnik, (1993). Ultrastructure of neuromuscular junctions in *Drosophila*: comparison of wild-type and mutants with altered excitability. J. Neurobiol. **24**, 1025-1044.

Johansen, J., M. Halpern and H. Keshishian, (1989). Axonal guidance and the development of muscle fiber-specific innervation in *Drosophila* embryos. J. Neurosci. **9**, 4318-4332.

Jungnickel, B., T. A. Rapoport and E. Hartmann, (1994). Protein translocation: common themes from bacteria to man. FEBS Letts **346**, 73-77.

Kaiser, K., (1990). From gene to phenotype in *Drosophila* and other organisms. BioEssays **12**, 297-301.

Kandler, K. and L. C. Katz, (1995). Neuronal coupling and uncoupling in the developing nervous system. Curr. Opin. Neurobiol. **5**, 98-105.

Kania, A., P. L. Han, Y. T. Kim and H. Bellen, (1993). *Neuromusculin* , a *Drosophila* gene expressed in peripheral neuronal precursors and muscles, encodes a cell-adhesion molecule. Neuron **11**, 673-687.

Karlin, S., G. Ghandour, F. Ost, S. Tavare and L. Korn, (1983). New approaches for computer analysis of nucleic acid sequences. Proc. Natl Acad. Sci. U. S. A. **80**, 5660.

Kennedy, T. E., T. Serafini, J. R. del Torre and M. Tessier-Lavigne, (1994). Netrins are diffusible chemotropic factors for commissural axons in the embryonic spinal cord. Cell **78**, 425-435.

Keshishian, H., A. Chiba, T. N. Chang, M. S. Halfon, E. W. Harkins, J. Jarecki, L. Wang, M. Anderson, S. Cash, M. E. Halpern and J. Johansen, (1993). Cellular mechanisms governing synaptic development in *Drosophila melanogaster*. J. Neurobiol. **24**, 757-787.

King, D. G. and R. Wyman, (1980). Anatomy of the giant fibre pathway in *Drosophila*. I. Three thoracic components of the pathway. J.Neurocytol. **9**, 753-770.

Kitagawa, H. and J. C. Paulson, (1994). Cloning of a novel α2,3-sialyltransferase that sialylates glycoprotein and glycolipid carbohydrate groups. J. Biol. Chem. **269**, 1394-1401.

Klämbt, C., J. R. Jacobs and C. S. Goodman, (1991). The midline of the *Drosophila* central nervous system: a model for the genetic analysis of cell fate, cell migration, and growth cone guidance. Cell **64**, 801-815.

Klapper, H. M., (1977). The independent distribution of amino acid near neighbour pairs into polypeptides. Biochem. Biophys. Res. Commun. **78**, 1018-1024.

Klein, P., M. Kanehisa and C. DeLisi, (1985). The detection and classification of membrane-spanning proteins. Biochim. Biophys. Acta **815**, 468-476.

Kolodkin, A., D. J. Matthes and C. S. Goodman, (1993). The *semaphorin* genes encode a family of transmembrane and secreted growth cone guidance molecules. Cell **75**, 1389-1399.

Kolodkin, A. L., D. J. Matthes, T. P. O'Connor, N. H. Patel, A. Admon, D. Bentley and C. S. Goodman, (1992). Fasciclin IV: Sequence, expression, and function during growth cone guidance in the grasshopper embryo. Neuron **9**, 831-845.

Koto, M., M. A. Tanouye, A. Ferrus, J. B. Thomas and R. J. Wyman, (1981). The morphology of the cervical giant fibre neuron of *Drosophila*. Brain Res. **221**, 213-217.

Kovalic, D., J.-H. Kwac and B. Weisblum, (1991). General method for direct cloning of DNA fragments generated by the polymerase chain reaction. Nucl. Acids Res. **19**, 4560.

Kozak, M., (1986). Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell **44**, 283-292.

Kozak, M., (1987). At least six nucleotides preceeding the AUG initiator codon enhance translation in mammalian cells. J. Mol. Biol. **196**, 947-950.

Kramers, P. G. N., A. P. Schalet, E. Paradi and L. Huiser-Hoogteyling, (1983). High proportion of multi-locus deletions among hycanthone-induced X-linked recessive lethals in *Drosophila melanogaster*. Mutat. Res. **107**, 187-201.

Krishnan, S. N., E. Frei, A. P. Schalet and R. J. Wyman, (1995). Molecular basis of intracistronic complementation in the *Passover* locus of *Drosophila*. Proc. Natl Acad. Sci. U. S. A. **92**, 2021-2025.

Krishnan, S. N., E. Frei, G. P. Swain and R. J. Wyman, (1993). *Passover*: a gene required for synaptic connectivity in the giant fibre system of *Drosophila*. Cell **73**, 967-977.

Kuwada, J. Y., (1986). Cell recognition by neuronal growth cones in a simple vertebrate embryo. Science **232**, 740-746.

Kyte, J. and R. F. Doolittle, (1982). A simple method for displaying the hydrophobic character of a protein. J. Mol. Biol. **157**, 105-132.

Laird, C. D. and B. J. McCarthy, (1968). Magnitude of interspecific nucleotide variation in *Drosophila*. Genetics **60**, 303-322.

Lefevre, G., (1981). The distribution of randomly recovered X-ray-induced sex-linked genetic effects in *Drosophila melanogaster*. Genetics **99**, 461-480.

Leptin, M., T. Bogaert, R. Lehmann and M. Wilcox, (1989). The function of PS integrins during *Drosophila* embryognesis. Cell **56**, 401-408.

Lesk, A. M., C.-I. Brandén and C. Chothia, (1989). Structural principles of alpha/beta barrel proteins: the packing of the interior of the sheet. Proteins **5**, 139-148.

Leung-Hagesteijn, C., A. M. Spence, B. D. Stern, Y. Zhou, M. W. Su, E. M. Hedgecock and J. G. Culotti, (1992). UNC-5, a transmembrane protein with immunoglobulin and thrombospondin type 1 domains, guides cell and pioneer axon migrations in *C. elegans*. Cell **71**, 289-299.

Lifschytz, E. and R. Falk, (1969). Fine structure analysis of a chromosome segment in *Drosophila melanogaster*. Analysis of ethyl methanesulphonate - induced lethals. Mutat. Res. **8**, 147-155.

Lifschytz, E. and N. Yakobovitz, (1978). The role of X-linked lethal and viable male sterile mutations in male gametogenesis of *Drosophila melanogaster*: Genetic analysis. Mol. Gen. Genet. **161**, 275-284.

Lin, D. M., V. J. Auld and C. S. Goodman, (1995). Targeted neuronal cell ablation in the *Drosophila* embryo- Pathfinding by follower growth cones in the absence of pioneers. Neuron **14**, 707-715.

Lindsley, D. L. and G. G. Zimm, (1992). The Genome of *Drosophila melanogaster*. (San Diego, CA : Academic Press, Inc).

Lipman, D. J. and W. R. Pearson, (1985). Rapid and sensitive protein similarity searches. Science **227**, 1435-1441.

Lipman, D. J., W. J. Wilbur, T. F. Smith and M. S. Waterman, (1984). On the statistical significance of nucleic acid similarities. Nucl. Acids Res. **12**, 215-226.

Lipp, J. and B. Dobberstein, (1986). The membrane-spanning segment of invariant chain (Iγ) contains a potentially cleavable signal sequence. Cell **46**, 1103-1112.

Lipp, J. and B. Dobberstein, (1988). Signal and membrane anchor function overlap in the type II membrane protein IγCAT. J. Cell. Biol. **106**, 1813-1820.

Lipshitz, H. D. and D. Kankel, (1985). Specificity of gene action during central nervous system development in *Drosophila melanogaster*: Analysis of the *lethal(1) optic ganglion reduced* locus. Dev. Biol. **108**, 56-77.

Lohmer, S., M. Maddaloni, M. Motto, F. Salamini and R. D. Thompson, (1993). Translation of the mRNA of the maize transcriptional activator *Opaque*-2 is inhibited by upstream open reading frames present in the leader sequence. Plant Cell **5**, 65-73.

Loughney, K., R. Kreber and B. Ganetzky, (1989). Molecular analysis of the *para* locus, a sodium channel gene in *Drosophila*. Cell **58**, 1143-1154.

Lumsden, A. G. S. and A. M. Davies, (1983). Earliest sensory nerve fibres are guided to peripheral targets by attractants other than nerve growth factor. Nature **306**, 786-788.

Lumsden, A. G. S. and A. M. Davies, (1986). Chemotropic effect of specific target epithelium in the developing mammalian nervous system. Nature **323**, 538-539.

Luo, Y., D. Raible and J. A. Raper, (1993). Collapsin: a protein in the brain that induces the collapse and paralysis of neuronal growth cones. Cell **75**, 217-227.

Luo, Y., I. Shepherd, J. Li, M. J. Renzi, S. Chang and J. A. Raper, (1995). A family of molecules related to collapsin in the embryonic chick nervous system. Neuron **14**, 1131-1140.

Matthes, D. J., H. Sink, A. L. Kolodkin and C. S. Goodman, (1995). Semaphorin II can function as a selective inhibitor of specific synaptic arborizations. Cell **81**, 631-639.

Matthews, R., (1992). A low-fat theory of anaesthesia. Science **255**, 156-157.

McCabe, P. C., (1990). Production of single stranded DNA by asymmetric PCR. In PCR Protocols: A Guide to Methods and Applications, M. A. Innis, D. H. Gelfand, J. J. Sninsky and T. J. White, eds. (San Diego: Academic Press), pp. 76-83.

McConnell, S. K., A. Ghosh and C. J. Shatz, (1989). Subplate neurons pioneer the first axon pathway away from the cerebral cortex. Science **245**, 978-982.

McGeoch, D. J., (1985). On the predictive recognition of signal peptide sequences. Virus Res. **3**, 271-286.

McIntire, S. L., G. Garriga, J. White, D. Jacobson and R. Horvitz, (1992). Genes necessary for directed axonal elongation or fasciculation in *C. elegans*. Neuron **8**, 307-322.

McKerracher, L., S. David, D. L. Jackson, V. Kottis, R. J. Dunn and P. E. Braun, (1994). Identification of myelin-associated glycoprotein as a major myelin-derived inhibitor of neurite growth. Neuron **13**, 805-811.

Mead, D. A., E. S. Skorupa and B. Kemper, (1985). Single stranded "Blue" T7 promoter plasmids: A versatile tandem promoter system for cloning and protein engineering. Nucl. Acids Res. **13**, 1103-1118.

Meier, T., S. Therianos, D. Zacharias and H. Reichert, (1993). Developmental expression of TERM-1 glycoprotein on growth cones and terminal arbors of individual identified neurons in the grasshopper. J. Neurosci. **13**, 1498-1510.

Merlie, J. P., D. Fagan, J. Mudd and P. Needleman, (1988). Isolation and characterisation of the complementary DNA for sheep seminal vesicle prostaglandin endoperoxide synthase (cyclooxygenase). J. Biol. Chem. **263**, 3550-3553.

Messersmith, E. K., E. D. Leonardo, C. J. Shatz, M. Tessier-Lavigne, C. S. Goodman and A. L. Kolodkin, (1995). Semaphorin III can function as a selective chemorepellent to pattern sensory projections in the spinal cord. Neuron **14**, 949-959.

Miletich, J. P. and G. J. Broze, (1990). β-protein C is not glycosylated at asparagine 329. J. Biol. Chem. **265**, 11397-11404.

Milks, L. C., N. M. Kumar, R. Houghten, N. Unwin and N. B. Gilula, (1988). Topology of the 32-kd liver gap junction protein determined by site-directed antibody localizations. EMBO J. **7**, 2967-2975.

Miklos, G. L. G., L. E. Kelly, P. E. Coombe, C. Leeds and G. Lefevre, (1987). Localisation of the genes *shaking-B*, *small optic lobes*, *sluggish-A*, *stoned*, and *stress-sensitive*-C to a well-defined region on the X- chromosome of *Drosophila melanogaster*. J. Neurogenet. **4**, 1-19.

Miklos, G. L. G., M. Yamamoto, J. A. Davies and V. Pirrotta, (1988). Microcloning reveals a high frequency of repetitive sequences characteristic of chromosome 4 and the β-heterochromatin of *Drosophila melanogaster*. Proc. Natl Acad. Sci. U. S. A. **85**, 2051-2055.

Montadon, A. J., P. M. Green, F. Gianelli and D. R. Bentley, (1989). Direct detection of point mutations: Application to haemophilia B. Nucl. Acids Res. **17**, 3347-3358.

Morgan, P. G., M. Sedensky and P. M. Meneely, (1990). Multiple sites of action of volatile anaesthetics in *Caenorhabditis elegans*. Proc. Natl Acad. Sci. U. S. A. **87**, 2965-2969.

Mount, S. M., (1982). A catalogue of splice junction sequences. Nucl. Acids Res. **10**, 459-472.

Mueller, M., I. Ibrahimi, C. N. Chang, P. Walter and G. Blobel, (1982). A bacterial secretory protein requires signal recognition particle for translocation across mammalian endoplasmic reticulum. J. Biol. Chem. **257**, 11860-11863.

Muesch, A., E. Hartmann, K. Rohde, A. Rubartelli, R. Sitia and T. A. Rapoport, (1990). A novel pathway for secretory proteins? TIBS **15**, 86-88.

Mukhopadhyay, G., P. Doherty, F. S. Walsh, P. R. Crocker and M. T. Fibin, (1994). A novel role for myelin assoociated glycoprotein as an inhibitor of axonal regeneration. Neuron **13**, 757-767.

Muralidhar, M. G. and J. B. Thomas, (1993). The *Drosophila bendless* gene encodes a neural protein related to ubiquitin-conjugating enzymes. Neuron **11**, 253-266.

Myers, R. M., S. G. Fischer, L. S. Lerman and T. Maniatis, (1985). Nearly all single base substitutions in DNA fragments joined to a GC clamp can be detected by denaturing gradient gel electrophoresis. Nucl. Acids Res. **13**, 3111-3130.

Nachtigall, W. and D. M. Wilson, (1967). Neuromuscular control of Dipteran flight. J. Exp. Biol. **47**, 77-97.

Nakashima, H. and K. Nishikawa, (1992). The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. FEBS Letts **303**, 141-146.

Ng, S. -C., L. A. Perkins, G. Conboy, N. Perrimon and M. C. Fishman (1989). A *Drosophila* gene expressed in the embryonic CNS shares one

conserved domain with mammalian GAP-43. Development **105**, 629-638.

Nose, A., V. B. Mahajan and C. S. Goodman, (1992). *Connectin*: A homophilic cell adhesion molecule expressed on a subset of muscles and the motoneurons that innervate them in *Drosophila*. Cell **70**, 553-567.

Nose, A., M. Takeichi and C. S. Goodman, (1994). Ectopic expression of Connectin reveals a repulsive function during growth cone guidance and synapse formation. Neuron **13**, 525-539.

Nothwehr, S. F., S. D. Hoeltzli, K. L. Allen, M. O. Lively and J. I. Gordon, (1990). Residues flanking the COOH-terminal C-region of a model eukaryotic signal peptide influence the site of its cleavage by signal peptidase and the extent of coupling of its co-translational translocation and proteolytic processing *in vitro*. J. Biol. Chem. **265**, 21797-21803.

Nozaki, Y. and C. Tanford, (1971). The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions. Establishment of a hydrophobicity scale. J. Biol. Chem. **246**, 2211-2217.

Oh, C. E., R. McMahon, S. Benzer and M. A. Tanouye, (1994). *bendless*, a *Drosophila* gene affecting neuronal connectivity, encodes a ubiquitin-conjugating enzyme homolog. J. Neurosci. **14**, 3166-3179.

O'Kane, C. J. and K. G. Moffat, (1992). Selective cell ablation and genetic surgery. Curr. Opin. Genet. Dev. **2**, 602-7.

Parks, G. D. and R. A. Lamb, (1993). Role of NH_2-terminal positively-charged residues in establishing membrane protein topology. J. Biol. Chem. **268**, 19101-19109.

Pastink, A., A. P. Schalet, C. Vreeken, E. Paradi and J. C. J. Eeken, (1987). The nature of radiation-induced mutations at the *white* locus of *Drosophila melanogaster*. Mutat. Res. **177**, 101-115.

Pastink, A., C. Vreeken, A. P. Schalet and J. C. J. Eeken, (1988). DNA sequence analysis of X-ray-induced deletions at the *white* locus of *Drosophila melanogaster*. Mutat. Res. **207**, 23-38.

Paul, C. and J. P. Rosenbusch, (1985). Folding patterns of Porin and Bacteriorhodopsin. EMBO J. **4**, 1593-1597.

Pearson, W. R. and D. J. Lipman, (1988). Improved tools for biological sequence data comparison. Proc. Natl Acad. Sci. U. S. A. **85**, 2444-2448.

Perlman, D. and H. O. Halvorson, (1983). A putative signal peptidase recognition site and sequence in eukaryotic and prokaryotic signal peptides. J. Mol. Biol. **167**, 391-409.

Perrimon, N. D., D. Smouse and G. L. G. Miklos, (1989). Developmental genetics of loci at the base of the X chromosome. Genetics **121**, 313-331.

Phelan, P., M. Nakagawa, M. B. Wilkin, Kevin G. Moffat, C. J. O'Kane, J. A. Davies and J. P. Bacon, (1996). Mutations in *shaking-B* prevent electrical synapse formation in the *Drosophila* giant fibre system. J. Neurosci., in Press.

Phillis, R. W., A. T. Bramlage, C. Wotus, A. Whittaker, L. S. Gramates, D. Seppala, F. Farahanchi, P. Caruccio and R. K. Murphey, (1993). Isolation of mutations affecting neural circuitry required for grooming behaviour in *Drosophila melanogaster*. Genetics **133**, 581-592.

Picot, D. and R. M. Garavito, (1994). Prostaglandin H synthase: Implications for membrane structure. FEBS Letts **346**, 21-25.

Picot, D., P. J. Loll and R. M. Garavito, (1994). The X-ray crystal structure of the membrane protein prostaglandin H_2 synthase-1. Nature **367**, 243-249.

Pinna, L. A., (1990). Casein kinase 2: an '*eminence grise*' in cellular regulation? Biochim. Biophys. Acta **1054**, 267-284.

Pirrotta, V., (1986). Cloning *Drosophila* genes. In *Drosophila*: A Practical Approach, D. B. Roberts, eds. (Oxford: IRL Press), pp. 83-110.

Ponce, M. R. and J. L. Micol, (1992). PCR amplification of long DNA fragments. Nucl. Acids Res. **20**, 623.

Poole, S. J., L. M. Kauvar, B. Drees and T. Kornberg, (1985). The *engrailed* locus of *Drosophila*: Structural analysis of an embryonic transcript. Cell **40**, 37-43.

Pringle, J. W. S., (1949). The excitation and contraction of the flight muscles of insects. J. Physiol. **108**, 226-232.

Proudfoot, N. J. and E. Whitelaw, (1988). Termination and 3' end processing of eukaryotic RNA. In Transcription and Splicing, B. D. Hames and D. M. Glover, eds. (Oxford: IRL Press), pp. 97-129.

Puschel, A. W., R. H. Adams and H. Betz, (1995). Murine Semaphorin-D Collapsin is a member of a diverse gene family and creates domains inhibitory for axonal extension. Neuron **14**, 941-948.

Rao, J. K. M. and P. Argos, (1986). A conformational preference parameter to predict helices in integral membrane proteins. Biochim. Biophys. Acta **869**, 197-214

Raper, J. A., M. J. Bastiani and C. S. Goodman, (1983). Pathfinding by neuronal growth cones in grasshopper embryos. II. Selective fasciculation onto specific axonal pathways. J. Neurosci. **3**, 31-41.

Raper, J. A. and J. P. Kapfhammer, (1990). The enrichment of a neuronal growth cone collapsing activity from embryonic chick brain. Neuron **4**, 21-29.

Rapoport, T., (1992). Transport of proteins across the endoplasmic reticulum membrane. Science **258**, 931-936.

Rapoport, T. A. and M. Wiedmann, (1985). Application of the signal hypothesis to the incorporation of integral membrane proteins. Curr. Top. Memb. Transp. **24**, 1-63.

Roeder, K. D., (1951). Movements of the thorax and potential changes in the thoracic muscles of insects during flight. Biological Bulletin **100**, 95-106.

Rose, G. D., A. R. Geselowitz, G. J. Lesser, R. H. Lee and M. H. Zehfus, (1985). Hydrophobicity of amino acid residues in globular proteins. Science **229**, 834-838.

Rosenblatt, M., N. V. Beaudette and G. D. Fasman, (1980). Conformational studies of the synthetic precursor-specific region of preproparathyroid hormone. Proc. Natl Acad. Sci. U. S. A. **77**, 3983-3987.

Roskies, A. L. and D. D. O'Leary, (1994). Control of topographic retinal axon branching by inhibitory membrane-bound molecules. Science **265**, 799-803.

Roy, P., C. Chatellard, G. Lemay, P. Crine and G. Boileau, (1993). Transformation of the signal peptide/membrane anchor domain of a type II transmembrane protein into a cleavable signal peptide. J. Biol. Chem. **268**, 2699-2704.

Rychlik, W. and R. E. Rhoads, (1989). A computer program for choosing optimal oligonucleotides for filter hybridisation, sequencing, and *in vitro* amplification of DNA. Nucl. Acids Res. **17**, 8543-8551.

Ryerse, J. S., (1989). Isolation and characterisation of gap junctions from *Drosophila melanogaster*. Cell Tissue Res. **256**, 7-16.

Ryerse, J. S., (1991). Gap junction protein tissue distribution and abundance in the adult brain in *Drosophila*. Tissue Cell **23**, 709-718.

Sambrook, J., E. F. Fritsch and T. Maniatis, (1989). Molecular Cloning: A Laboratory Manual. (New York : Cold Spring Harbor Laboratory Press).

Sankhoff, D. and R. Cedergren, (1973). A test for nucleotide sequence homology. J. Mol. Biol. **77**, 159.

Schalet, A. and G. Lefevre, (1976). The proximal region of the X chromosome. In The Genetics and Biology of *Drosophila*, M. Ashburner and E. Novitski, eds. (New York: Academic Press), pp. 845-902.

Schildkraut, C. and S. Lifson, (1965). Dependence of the melting temperature of DNA on salt concentration. Biopolymers **3**, 195-208.

Schmid, S. R. and M. Spiess, (1988). Deletion of the amino-terminal domain of asialoglycoprotein receptor H1 allows cleavage of the internal signal sequence. J. Biol. Chem. **263**, 16886-16891.

Sedensky, M. M. and P. M. Meneely, (1987). Genetic analysis of halothane sensitivity in *Caenorhabditis elegans*. Science **236**, 952-954.

Serafini, T., T. Kennedy, M. Galko, C. Mirzyan, T. Jessell and M. Tessier-Lavigne, (1994). The Netrins define a family of axon outgrowth-promoting proteins with homology to *C. elegans UNC-6*. Cell **78**, 409-424.

Sharp, P. M., E. Cowe, D. G. Higgins, D. C. Shields, K. H. Wolfe and F. Wright, (1988). Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*: a review of the considerable within-species diversity. Nucl. Acids Res. **16**, 8207-8211.

Shaw, A. S., P. M. Rottier and J. K. Rose, (1988). Evidence for the loop model of signal sequence insertion into the endoplasmic reticulum. Proc. Natl Acad. Sci. U. S. A. **85**, 7592-7596.

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

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

Shirasaki, R., A. Tamada, R. Katsumata and F. Murakami, (1995). Guidance of cerebellofugal axons in the rat embryo- Directed growth toward the floor plate and subsequent elongation along the longitudinal axis. Neuron **14**, 961-972.

Short, J. M., J. M. Fernandez, J. A. Sorge and W. Huse, (1988). λZAP: An expression vector with *in vitro* excision properties. Nucl. Acids Res. **16**, 7583-7599.

Shugihara, H., V. Andrisani and P. M. Salvaterra, (1990). *Drosophila* choline acetyl transferase uses a non-AUG initiation codon and full length RNA is inefficiently translated. J. Biol. Chem. **265**, 21714-21719.

Singh, N. R., K. Singh and D. R. Kankel, (1989). Development and fine structure of the nervous system of *lethal(1)optic ganglion reduced* visual mutants of *Drosophila melanogaster*. In Neurobiology of Sensory Systems, R. N. Singh and N. J. Strausfeld, eds. (New York: Plenum Press).

Sink, H. and P. M. Whitington, (1991a). Early ablation of target muscles modulates the arborization pattern of an identified embryonic *Drosophila* motoneuron. Development **113**, 701-707.

Sink, H. and P. M. Whitington, (1991b). Location and connectivity of abdominal motoneurons in the embryo and larva of *Drosophila melanogaster*. J. Neurobiol. **22**, 298-311.

Sipos, L. and G. von Heijne, (1993). Predicting the topology of eukaryotic membrane proteins. Eur. J. Biochem. **213**, 1333-1340.

Smith, G. E. and M. D. Summers, (1980). The bidirectional transfer of DNA and RNA to nitrocellulose or diazobenzyloxymethyl paper. Anal. Biochem. **109**, 123.

Snow, E. T., R. S. Foote and S. Mitra, (1984). Base-pairing properties of O^6-methylguanine in template DNA during *in vitro* replication. J. Biol. Chem. **259**, 8095-8100.

Snow, P. M., A. J. Bieber and C. S. Goodman, (1989). Fasciclin III: A novel homophilic adhesion molecule in *Drosophila*. Cell **59**, 313-323.

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

Staden, R., (1984). Computer methods to locate signals in nucleic acid sequences. Nucl. Acids Res. **12**, 505-519.

Staden, R. and A. D. McLachlan, (1982). Codon preference and its use in identifying protein coding regions in long DNA sequences. Nucl. Acids Res. **10**, 141-156.

Starich, T. A., R. K. Herman and J. E. Shaw, (1993). Molecular and genetic analysis of *unc-7*, a *Caenorhabditis elegans* gene required for coordinated locomotion. Genetics **133**, 527-541.

Steele, J., (1982). J. Appl. Math. **42**, 731.

Strausfeld, N. J. and M. Obermeyer, (1976). Resolution of intraneuronal and transsynaptic migration of cobalt in the insect visual system. J. Comp. Physiol. (A) **110**, 1-12.

Sturtevant, A. H., (1920). Genetic studies on *Drosophila simulans*. 1. Introduction. Hybrids with *Drosophila melanogaster*. Genetics **5**, 488-500.

Suchyna, T. M., L.-X. Xu, F. Gao, C. R. Fourtner and B. J. Nicholson, (1993). Identification of a proline residue as a transduction element involved in voltage gating of gap junctions. Nature **365**, 847-849.

Swain, G. P., R. J. Wyman and M. D. Egger, (1990). A deficiency chromosome in *Drosophila* alters neuritic projections in an identified motoneuron. Brain Res. **535**, 147-150.

Swenson, K. I., J. R. Jordan, E. C. Beyer and D. L. Paul, (1989). Formation of gap junctions by expression of Connexins in *Xenopus* oocyte pairs. Cell **57**, 145-155.

Szczesna-Skorupa, E. and B. Kemper, (1989). NH_2-terminal substitutions of basic amino acids induce translocation across the microsomal membrane and glycosylation of rabbit cytochrome P450IIC2. J. Cell Biol. **108**, 1237-1243.

Taghert, P. H., M. J. Bastiani, R. K. Ho and C. S. Goodman, (1982). Guidance of pioneer growth cones: Filopodial contacts and coupling revealed with an antibody to Lucifer Yellow. Dev. Biol. **94**, 391-399.

Tamada, A., R. Shirasaki and F. Murakami, (1995). Floor plate chemoattracts crossed and chemorepulses uncrossed axons in the vertebrate brain. Neuron **14**, 1083-1093.

Tang, J., L. Landmesser and U. Rutishauser, (1992). Polysialic acid influences specific pathfinding by avian motoneurons. Neuron **8**, 1031-1044.

Tanouye, M. A. and D. G. King, (1983). Giant fibre activation of direct flight muscles in *Drosophila*. J. Exp. Biol. **105**, 241-251.

Tanouye, M. A. and R. J. Wyman, (1980). Motor outputs of the giant nerve fiber in *Drosophila*. J. Neurophysiol. **44**, 405-421.

Tepass, U. and V. Hartenstein, (1994). The development of cellular junctions in the *Drosophila* embryo. Dev. Biol. **161**, 563-596.

Tessier-Lavigne, M., (1994). Axon guidance by diffusible repellents and attractants. Curr. Opin. Genet. Dev. **4**, 596-601.

Thomas, J. B. and R. J. Wyman, (1982). A mutation in *Drosophila* alters connectivity between two identified neurones. Nature **298**, 650-651.

Thomas, J. B. and R. J. Wyman, (1984a). Duplicated neural structure in *bithorax* mutants of *Drosophila*. Dev. Biol. **102**, 531-533.

Thomas, J. B. and R. J. Wyman, (1984b). Mutations altering synaptic connectivity between identified neurons in *Drosophila*. J. Neurosci. **4**, 530-538.

Thomas, J. H., (1993). Thinking about genetic redundancy. TIG **9**, 395-399.

Tibbitts, T. T., D. L. D. Caspar, W. C. Phillips and D. A. Goodenough, (1990). Diffraction diagnosis of protein folding in gap junction connexons. Biophys. J. **57**, 1025-1036.

Twitty, V. C., (1937). Experiments on the phenomenon of paralysis produced by a toxin ocurring in *Trituris* embryos. J. Exp. Zool. **76**, 67-104.

Twitty, V. C. and H. A. Elliot, (1934). Motor inhibition in *Amblystoma* produced by *Trituris* transplants. Science **80**, 78-79.

Unwin, N., (1989). The structure of ion channels in membranes of excitable cells. Neuron **3**, 665-676.

Van Vactor, D., H. Sink, D. Fambrough, R. Tsoo and C. S. Goodman, (1993). Genes that control neuromuscular specificity in *Drosophila*. Cell **73**, 1137-1153.

Verselis, V., M. V. L. Bennett and T. A. Bargiello, (1991). A voltage dependent gap junction in *Drosophila melanogaster*. Biophys. J. **59**, 114-126.

Vogel, H., J. K. Wright and F. Jähnig, (1985). The structure of lactose permease derived from Raman spectroscopy and prediction methods. EMBO J. **4**, 3625-3631.

von Heijne, G., (1985). Signal sequences: the limits of variation. J. Mol. Biol. **184**, 99-105.

von Heijne, G., (1986). A new method for predicting signal sequence cleavage sites. Nucl. Acids Res. **14**, 4683-4691.

von Heijne, G., (1990). The signal peptide. J. Membrane Biol. **115**, 195-201.

von Heijne, G., (1991). Proline kinks in transmembrane α helices. J. Mol. Biol. **218**, 499-503.

von Heijne, G., (1992). Membrane protein structure predicition: Hydrophobicity analysis and the positive inside rule. J. Mol. Biol. **225**, 487-494.

von Heijne, G. and Y. Gavel, (1988). Topogenic signals in integral membrane proteins. Eur. J. Biochem. **174**, 671-678.

Wallace, J. C. and S. Henikoff, (1992). PATMAT: A searching and extraction program for sequence, pattern and blocks queries and databases. CABIOS **8**, 249-254.

Wallace, R. B. and C. G. Miyada, (1987). Oligonucleotide probes for the screening of recombinant DNA libraries. In Guide to Molecular Cloning Techniques, S. L. Berger and A. R. Kimmel, eds. (San Diego: Academic Press), pp. 432-442.

Watanabe, T. and D. R. Kankel, (1990). Molecular cloning and analysis of *l(1)ogre*, a locus of *Drosophila melanogaster* with prominent effects on the postembryonic development of the central nervous system. Genetics **126**, 1033-1044.

Watanabe, T. and D. R. Kankel, (1992). The *l(1)ogre* gene of *Drosophila melanogster* is expressed in postembryonic neuroblasts. Dev. Biol. **152**, 172-183.

Waterston, R. H., (1988). Muscle. In The Nematode *Caenorhabditis elegans*, W. B. Wood, ed. (Cold Spring Harbor, N. Y.: Cold Spring Harbor Laboratory Press).

Wehrle, B. and M. Chiquet, (1990). Tenascin is accumulated along developing peripheral nerves and allows neurite outgrowth *in vitro*. Development **110**, 401-415.

Weinzierl, R., J. M. Axton, A. Ghysen and M. Akam, (1987). *Ultrabithorax* mutations in constant and variable regions of the protein coding sequence. Genes Dev. **1**, 386-397.

Wessels, H. P. and M. Spiess, (1988). Insertion of a multispanning membrane protein occurs sequentially and requires only one signal sequence. Cell **55**, 61-70.

Westerfield, M., D. W. Liu, C. B. Kimmel and C. Walker, (1990). Pathfinding and synapse formation in a zebrafish mutant lacking functional acetylcholine receptors. Neuron **4**, 867-874.

Wilson-Rawls, J., S. L. Deutscher and W. S. M. Wold, (1994). The signal-anchor domain of adenovirus E3-6.7K, a type III integral membrane protein, can direct adenovirus E3-gp19K, a type I integral membrane protein, into the endoplasmic reticulum. Virology **201**, 66-76.

Woodget, J. R., K. L. Gould and T. Hunter, (1986). Substrate specificity of protein kinase C. Use of synthetic peptides corresponding to physiological sites as probes for substrate recognition requirements. Eur. J. Biochem. **161**, 177-184.

Wu, C. -F., B. Ganetzky, Y. N. Jan and L. Y. Jan, (1978). A *Drosophila* mutant with a temperature sensitive block in nerve conduction. Proc. Natl Acad. Sci. U. S. A. **75**, 4047-4051.

Wyman, R. J. and J. B. Thomas, (1983). What genes are necessary to make an identified synapse ? C. S. H. S. Q. B. **48**, 641-652.

Yamamoto, M.-T. and G. Miklos, (1987). Cytological analysis of *Deficiency 16-3-35* at the base of the X chromosome of *Drosophila melanogaster*. Drosophila Information Service **66**, 154.

Yancey, S. B., S. A. John, R. Lal, B. J. Austin and J. Revel, (1989). The 43-kD polypeptide of heart gap junctions: immunolocalisation, topology, and functional domains. J. Cell Biol. **108**, 2241-2254.

Yu, J., J.-H. Lin, X.-R. Wu and T.-T. Sun, (1994). Uroplakins Ia and Ib, two major differentiation products of bladder epithelium, belong to a family of four transmembrane domain (4TM) proteins. J. Cell Biol. **125**, 171-182.

Zerial, M., D. Huylenbroeck and H. Garoff, (1987). Foreign transmembrane peptides replacing the internal signal sequence of transferrin receptor allow its translocation and membrane binding. Cell **48**, 147-155.

Zhang, H., R. Scholl, J. Browse and C. Somerville, (1988). Double stranded sequencing as a choice for DNA sequencing. Nucl. Acids Res. **16**, 1220.

Zheng, J. Q., M. Felder, J. A. Connor and M.-m. Poo, (1994). Turning of nerve growth cones induced by neurotransmitters. Nature **368**, 140-144.

Zhong, Y., V. Budnik and C. -F. Wu, (1992). Synaptic plasticity in *Drosophila* memory and hyperexcitable mutants: role of the cAMP cascade. J. Neurosci. **12**, 644-651.

Zuker, M. and P. Stiegler, (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucl. Acids Res. **9**, 133.