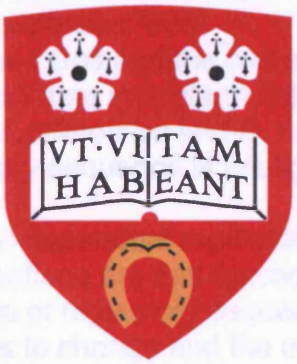
The Evolution of Bicoid Interactions in the Higher Diptera

Thesis submitted for the degree of Doctor of Philosophy
at the University of Leicester



by
Naomi S Wratten (B.Sc.)
Department of Genetics
University of Leicester

September 2003

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

The Evolution of Bicoid Interactions in the Higher Diptera

Naomi S Wratten

Development can be described as a network of genetic interactions. These interactions can be highly conserved over large evolutionary distances or they can vary between closely related species resulting in morphological differences. What are the forces promoting change in such interactions and how do they evolve?

The interaction between the transcription factor Bicoid (Bcd) and the *hunchback* (*hb*) promoter was compared between *M. domestica* and *D. melanogaster* (Bonneton *et al.*, 1997; Shaw *et al.*, 2002). This interaction is conserved in function despite differences in both the Bcd homeodomain and the *hb* promoter sequences. Functional tests of the components of this interaction suggest that they are co-evolving in each species to maintain function in spite of sequence divergence.

In this thesis, two further Bcd regulated genes, *tailless* (*tll*) and *caudal* (*cad*), were studied to provide a comparison to the Bcd-*hb* promoter interaction and to investigate the consequences of regulatory sequence change within a network of interactions. The *tll* promoter sequences are unalignable between *M. domestica* and *D. melanogaster* yet are similar in function. As with the Bcd-*hb* promoter interaction functional tests indicate that the interaction is diverging between the species at the molecular level.

The interaction between Bcd and the *cad* mRNA was also investigated. *cad* sequence and expression data indicate that the function and regulation of *cad* is conserved between *M. domestica* and *D. melanogaster*. However, the *M. domestica cad* 3' regulatory sequence is unalignable with that of *D. melanogaster*.

In conclusion, the Bcd-dependent regulatory sequences are evolving relatively quickly but all indications are that function is conserved. This raises questions about the structure of regulatory sequences, the flexibility of non-coding regulatory sequences to change and the evolution of interactions and of non-coding regions in general.

Acknowledgements

I wish to thank the following people for their help over the last three years or more. Firstly my supervisor Gabby for having an evolutionary explanation for every strange result and for saying exactly what he thinks. To Phil Shaw for advice with experiments, constant encouragement and interesting scientific discussion.

John Hancock for analysing my data for Chapter 5. The people who kindly donated flies and Michael Stauber for also being interested in Bcd. Fred for his advice on experiments and life and Mark for introducing me to Freehand.

Thanks to the Genetics department and all the interesting characters who helped to make Leicester a fun and friendly place to work, including Pat the 'fit' gran, Ben the 'cool' guy and Zoë and her obsession with order. There are many others who I won't mention but who I will remember fondly!!

I would like to thank my parents for having their own life, forgetting to phone me and for being great friends. Rich for being the black sheep and making anything I did look mild in comparison and Jo for telling me what to do. Finally I would like to thank my newly acquired husband, Alistair, for without him I would still be trying to clone *caudal* and write my first Chapter! Thanks for the patience, support and laughter.

Communications

Some of the work carried out for this thesis was published in the following article:

Shaw, P. J., Wratten, N. S., McGregor, A. P. and Dover, G. A. (2002) Co-evolution in *bicoid*-dependent promoters and the inception of regulatory incompatibilities among species of higher Diptera. *Evol Dev* **4**:265-277.

List of Abbreviations

A	adenine
BSA	bovine serum albumin
C	cytosine
cDNA	complementary DNA
DEPC	diethylpyrocarbonate
DNaseI	deoxyribonuclease I
DTT	dithiothreitol
EDTA	ethylenediaminetetraacetic acid
G	guanine
GST	glutathione-S-transferase
HA	hemagglutinin
MY	million years
MYA	million years ago
OD	optical density
ORF	Open reading frame
PAGE	polyacrylamide gel electrophoresis
PNK	polynucleotide kinase
PNS	Peripheral nervous system
RNAi	RNA interference
SSC	sodium saline citrate
T	thymine
TAE	Tris-acetate EDTA
TE	Tris EDTA
TBE	Tris-borate EDTA
UTR	untranslated region

Contents

Chapter 1 General Introduction: The Evolution of Bicoid Interactions in the Higher Diptera

1.1 The evolution of development

In the last twenty years our knowledge of the molecular basis of development has vastly increased. We now know that many of the genes involved in development are phylogenetically conserved. For example, the *Hox* genes, which were originally discovered in *D. melanogaster,* have subsequently been found in all animal phyla (Lewis, 1978; for review see Prince and Pickett 2002). If the same genes are used in development in distinct taxa, how do different morphologies arise?

The answer lies in the redeployment of these same genes to different roles in the development of different body plans. This means that the position in the developmental network that each genes occupies and its range of interactions varies between species. The apparent flexibility of these genetic interactions is due in part to many of the genes encoding transcription factors. Thus part of the difference in developmental programs is a result of changes in the regulatory network of conserved genes (Carroll *et al.*, 2001).

1.2 Evolution of conserved genes

The basis of much of our understanding of the diversification of body plans came from studies of the *Hox* genes. Throughout animal evolution these genes have been repeatedly duplicated such that *D. melanogaster* has two clusters of *Hox* genes, mice have four and zebrafish have seven (Holland *et al.*, 1994, Amores *et al.*, 1998). Duplicated *Hox* genes have evolved new functions (Davis and Capecchi, 1996; Zakany and Duboule, 1999), yet often the coding regions are functionally interchangeable. For example, in mice the paralogues *Hoxa3* and *Hoxd3* have different roles in development yet a mouse with the *Hox3a* coding region inserted into the *Hoxd3* locus is completely viable (Greer *et al.*, 2000; Krumlauf, 1994; Akam, 1994). This divergence of function has been linked to changes within the regulatory regions of these genes. For example, *Hoxc8* is expressed in overlapping but different domains within the

neural tube, between mouse, chick and baleen whale. These differences in expression have been mapped to changes in transcription factor binding sites within the *Hoxc8* regulatory regions and are thought to directly affect the number of thoracic vertebrae in each species (Belting *et al.*, 1998; Shashikant *et al.*, 1998). Not all changes in development arise from divergence of *cis*-regulatory modules but it is accepted that part of the evolution of developmental programs is a result of changes in regulatory interactions of conserved genes (Galant and Carroll, 2002; Ronshaugen *et al.*, 2002; Carroll *et al.*, 2001).

It has been proposed that the functional redundancy created by duplications (such as in the *Hox* genes) enabled the divergence of paralogous genes to new functions. This divergence could involve subfunctionalisation of the paralogues and can explain the retention of large numbers of duplicated genes within the genome (Lynch and Force, 2000). Evidence for the evolution of regulatory regions is now being found in genes other than the *Hox* genes and without an *a priori* duplication event. For example, the pattern of hairs on the adult second leg and the larval body differ between closely related species of *Drosophila* due to differences in the regulatory regions of the genes *Ultrabithorax (Ubx)* and *shaven-baby (svb)* (Stern, 1998; Sucena and Stern, 2000; Sucena *et al.*, 2003).

Changes in transcription factors or within regulatory modules may result in the establishment of a new regulatory interaction. If the newly regulated gene is also a transcription factor then the downstream targets of this gene can also be co-opted to a new position in the developmental network. The evolution of butterfly wing spots is a result of co-option of the hedgehog signaling pathway from early wing development (Keys *et al*, 1999). Similarly, the vertebrate *Hox* genes were primarily involved in specification of the body segments but through co-option of the regulatory network the same genes are utilized later in development in specification of the developing limbs (Zakany *et al.*, 1997).

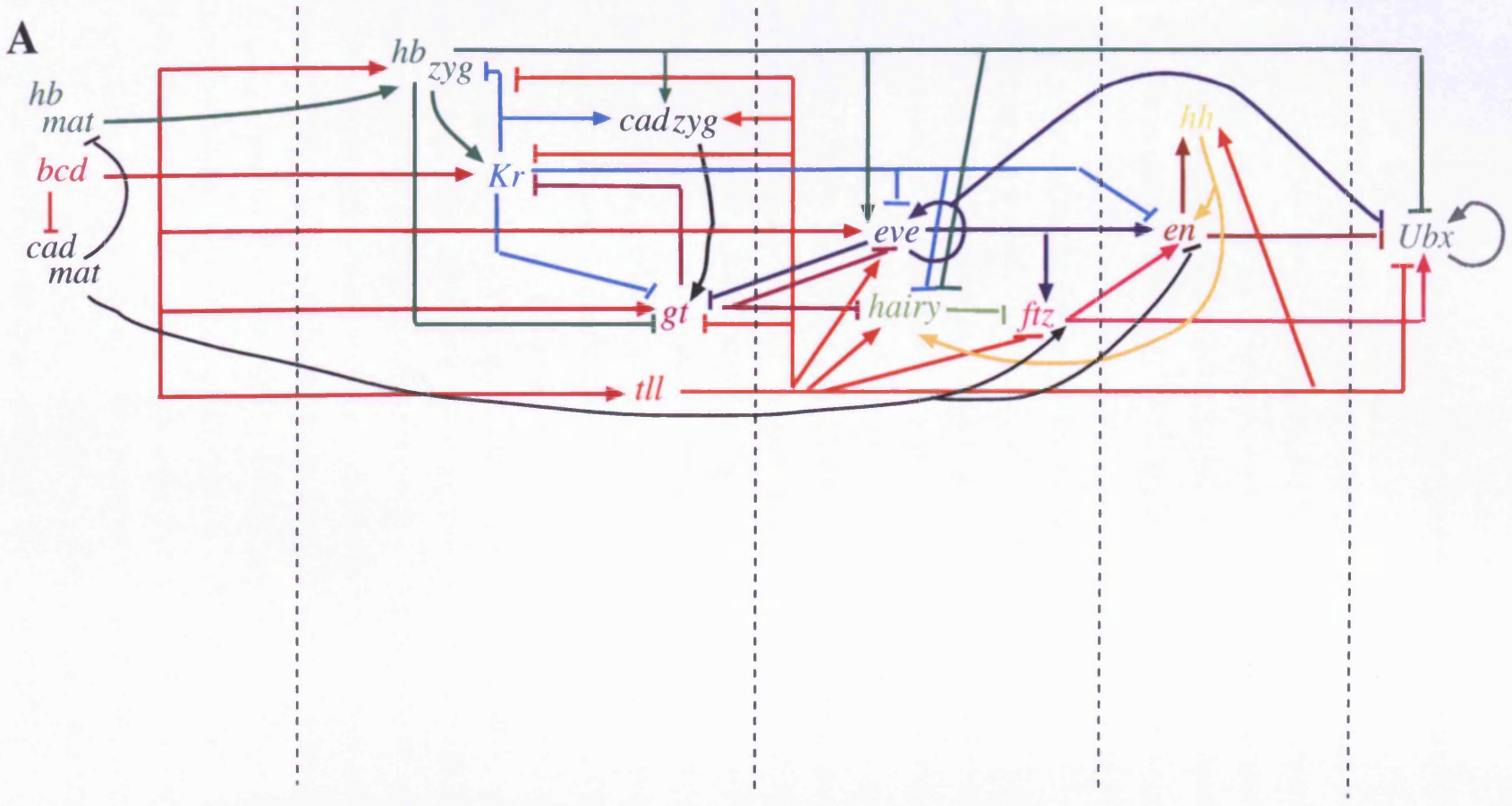
1.3 Features of developmental gene *cis*-regulatory sequences

The early stages of *Drosophila* development can be viewed as a network of genetic interactions (for example, see fig. 1.1A). This begins with a few maternally provided transcription factors within the egg and develops in complexity as genes are activated within the embryonic nuclei. Physically the egg progresses from an unspecified syncytium to a highly patterned embryo during this process (see fig. 1.1B; Lawrence, 1992).

Developmental genes typically have multiple functions, which can be seen as a complex pattern of expression at different developmental time points and in different tissues. To produce such expression patterns the cis-regulatory regions respond to many different regulatory cues. These regions vary in size with some of the most complex patterns such as *Ubx* being driven by regulatory regions spanning 70 kb (Martin *et al.*, 1995). The regulatory regions can contain binding sites for multiple regulatory factors, some of which interact with themselves and/or other factors, this means that spacing between binding sites can be important for function. The output of the promoter is the result of all the different positive and negative regulatory inputs acting on it at a particular time and place. Some well studied developmental gene promoters are shown in fig. 1.2.

The regulatory regions are made up of discrete blocks of sequence called modules. Each module drives part of the expression pattern of the gene. Thus, it is possible for one module to evolve whilst the other modules remain conserved. A module such as *eve* stripe 2 is only one of many regulatory regions responsive to Bicoid (Bcd), Hunchback (Hb), Giant (Gt) and Kruppel (Kr; see fig. 1.2A). Therefore, a module can be envisaged as part of a set of cis-regulatory elements all responding in part to a particular transcription factor (Davidson, 2001). As a result a change in the expression or function of a transcription factor can have extensive knock–on effects on downstream genetic pathways (Davidson, 2001).

Evidence from transgenic experiments has shown that the multiple


A
hb
mat
bcd
cad
mat
hb
zyg
cad
zyg
Kr
gt
tll
eve
hairy
ftz
hh
en
Ubx

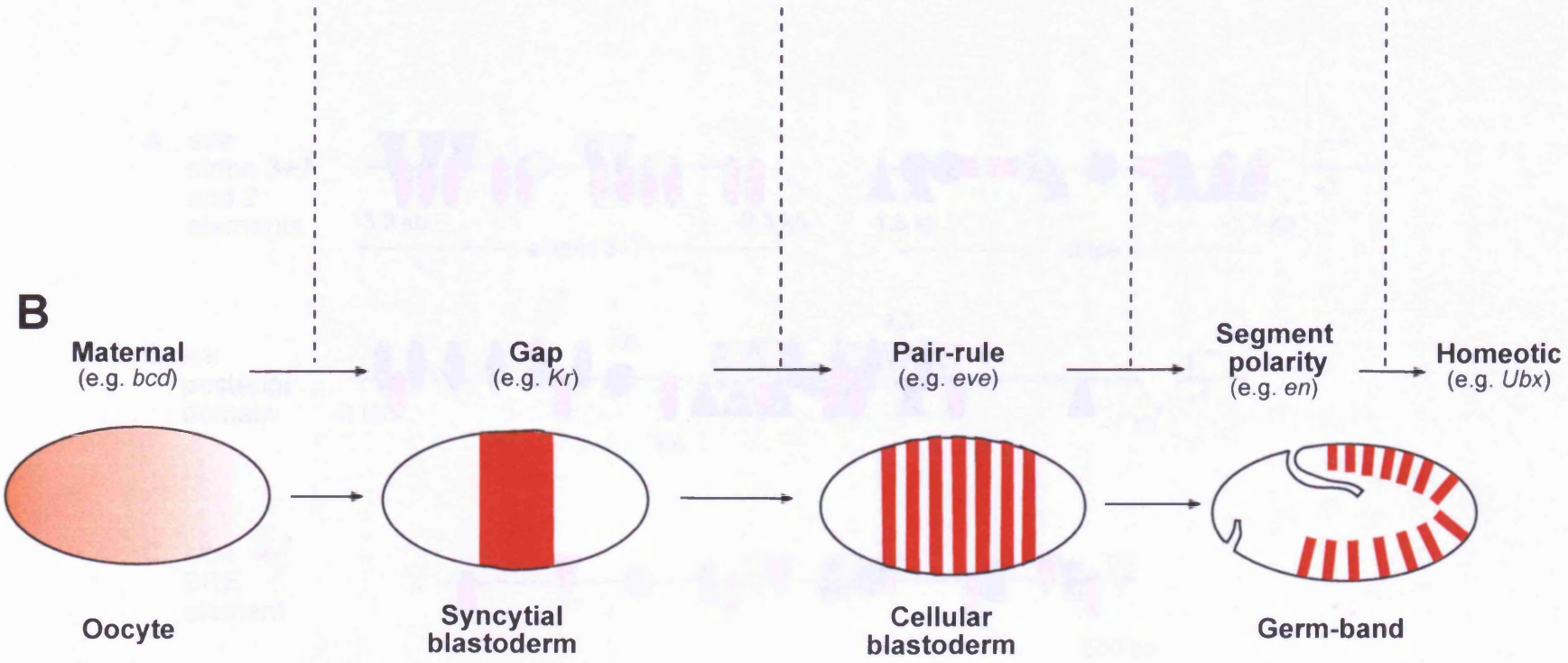


Figure 1.1A The gene regulatory network of early *Drosophila* development (**A**) and the early stages in the development of the *Drosophila* embryo (**B**)
A. Partial network of maternal (mat) and zygotic (zyg) interactions controlling anterior-posterior development in *Drosophila*. Arrowheads indicate activation and truncated lines represent repression (not all interactions are shown). Dashed vertical lines delineate the developmental classes to which each gene belongs. The classes are listed below in 1.1B (adapted from Sauer *et al.*, 1996).
B. Four major stages of *Drosophila* embryonic development are shown with development proceeding from left to right. Embryos are orientated with the anterior to the left and dorsal up. The protein distribution represented by red shading is typical of the genes expressed at that stage. The stages are shown in relation to the developmental gene hierarchy seen in 1.1A and an example of each class of gene is also listed.

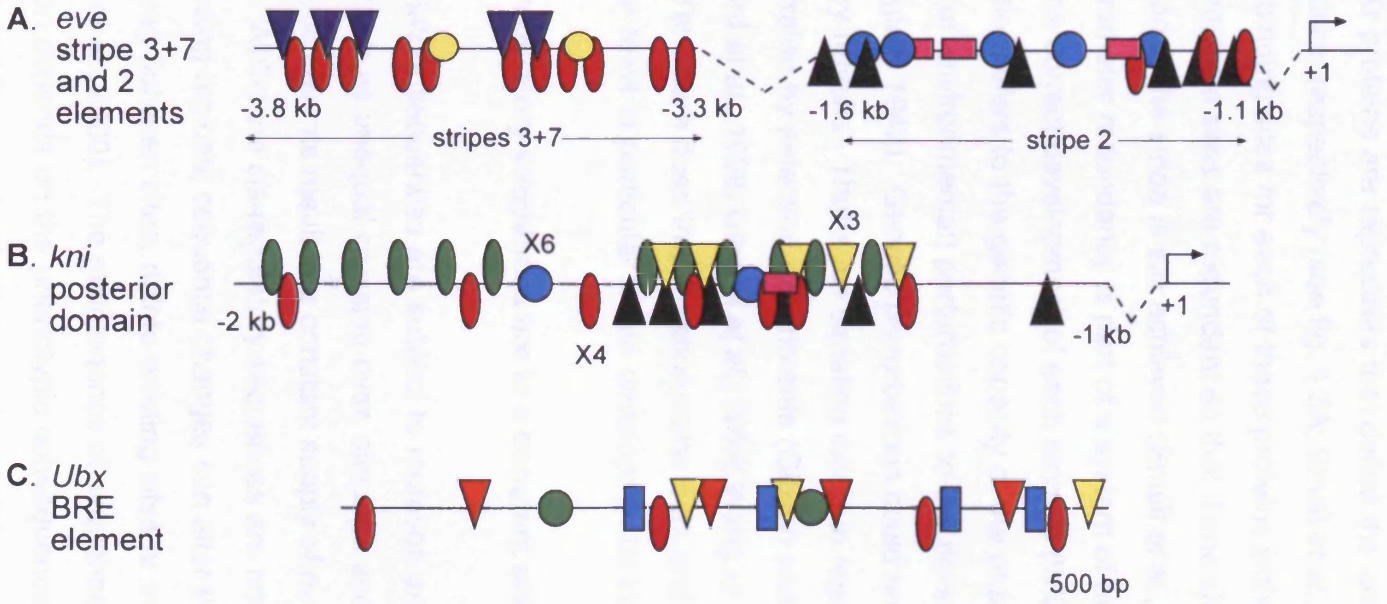


Figure 1.2 *Cis*-regulatory modules responsible for the expression of the *Drosophila* segmentation genes *eve* (**A**), *kni* (**B**) and *Ubx* (**C**) Each shape represents an individual binding site for the following transcription factors: Hunchback (red ovals), Bicoid (blue circles), Tailless (yellow triangles), Krüppel (black triangles), Caudal (green ovals), Giant (pink rectangles), Knirps (yellow circles) Twist (orange triangles), Engrailed (green circles), Fushi tarazu (blue rectangles) and D-Stat (blue triangles). Multiple binding sites are labelled with the number of sites present, for example X6. The scale in **A** and **B** is given in kb upstream from the transcription start site, which is represented by black arrows positioned at +1. In **A**, the two independent modules are indicated by the horizontal arrows and are labeled according to the stripes of *eve* expression they generate.
Parts **A** and **B** adapted from Arnone and Davidson 1997, figure 1. The *eve* promoter was characterised by Small *et al.*, 1992, 1996 and the *kni* promoter by Rivera-Pomar *et al.*, 1995. The *Ubx* BRE element drives expression in parasegments 6, 8, 10 and 12 (Qian *et al.*, 1993; Zhang *et al.*, 1991).

binding sites within a module can function redundantly. For example within the *eve* stripe 2 module both Bcd and Hb proteins are activating expression whilst Gt and Kr proteins are repressors that define the anterior and posterior borders of expression respectively (see fig. 1.3A; Small *et al.*, 1991, 1992). There are multiple binding sites for each of these proteins within the module. It is thought that the multiple sites are redundant so that if one site is altered the correct expression of the stripe is still achieved (Small *et al.*, 1992; Arnosti *et al.*, 1996).

Promoter redundancy is part of a system of buffering which helps to enable the correct development of each embryo (Wilkins, 1997). Buffering or canalisation refers to the genetic capacity of the organism to protect against genetic (or environmental) perturbations to the developmental process (Waddington, 1942). Genetic perturbations could result from polymorphic *cis*-regulatory regions. That such variation exists in regulatory regions has been demonstrated by selection experiments (Gibson and Van Helden, 1997; Rutherford *et al.*, 1998; Gibson *et al.*, 1999; Wang *et al.*, 1999; Robin *et al.*, 2002). Yet where does this variation come from and how is it tolerated at the molecular level in particular within developmental interactions?

1.4 *Cis*-regulatory sequences are in a constant state of flux

Cis-regulatory sequences are subject to mutation and genomic turnover events, such as unequal crossing over, slippage and gene conversion (Dover, 1993). Such events result in a constant supply of new *cis*-regulatory sequence variants. Whilst the *cis*-regulatory sequences are not constrained to the extent of the coding regions; sequence changes can alter the affinity of a binding site, the spacing between sites, delete existing sites or even create new ones (Ludwig *et al.*, 2000). The maintenance of sequence variants within the population depends on the phenotypic consequences of the sequence change. It is expected that most selection on *cis*-regulatory sequences will be to maintain the correct output of gene expression. Indeed a number of comparisons of homologous *cis*-regulatory modules show that function can be

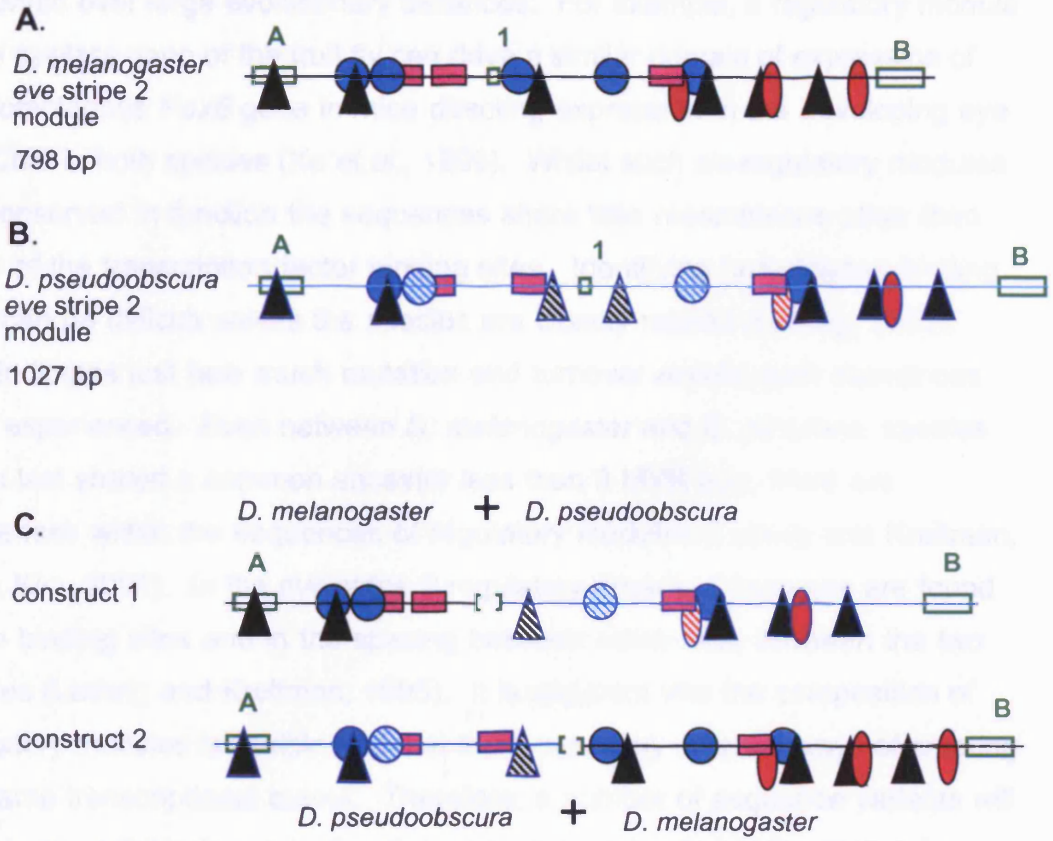


Figure 1.3 The structure of the *eve* stripe 2 module in *D. melanogaster* (**A**) and *D. pseudoobscura* (**B**); and an inter-specific test of *eve* stripe 2 function (**C**). Each shape represents an individual binding site for the following transcription factors: Hunchback (red ovals), Bicoid (blue circles), Krüppel (black triangles) and Giant (pink rectangles). Green boxes A and B indicate the 5' and 3' ends of the module and box 1 indicates a stretch of seven bases present in both *D. melanogaster* and *D. pseudoobscura*. The module lengths are given in bp in **A** and **B**. To differentiate between the two species the *D. melanogaster* module is outlined in black (**A**) and the *D. pseudoobscura* module in blue (**B**). Striped binding sites indicate sequences with a poor match to the consensus binding site sequence.
C. Two chimaeric constructs were made to test the function of the *eve* stripe 2 promoter in both species. The conserved sequence in box 1 was used to join the 5' half of one species *eve* stripe 2 module to the 3' half of the other species. The chimaeras were placed upstream of a *lacZ* reporter and tested in transgenic *D. melanogaster* (see text). The *D. melanogaster* and *D. pseudoobscura eve* stripe 2 modules were characterised by Small *et al.*, 1992, 1996 and Ludwig and Kreitman 1998, respectively. The chimaeras were created and tested by Ludwig *et al.*, 2000.

conserved over large evolutionary distances. For example, a regulatory module of the *eyeless* gene of the fruit fly can drive a similar domain of expression of the homologous *Pax6* gene in mice directing expression in the developing eye and CNS in both species (Xu *et al.*, 1999). Whilst such *cis*-regulatory modules are conserved in function the sequences share little resemblance other than those of the transcription factor binding sites. Identifying homologous binding sites can be difficult unless the species are closely related (Ludwig, 2002). This indicates just how much mutation and turnover events such sequences have experienced. Even between *D. melanogaster* and *D. simulans*, species which last shared a common ancestor less than 3 MYR ago, there are differences within the sequences of regulatory modules (Ludwig and Kreitman, 1995; Kim, 2001). In the *eve* stripe 2 regulatory module differences are found within binding sites and in the spacing between some sites between the two species (Ludwig and Kreitman, 1995). It is apparent that the composition of regulatory modules is flexible and that there are many different ways of creating the same transcriptional output. Therefore, a number of sequence variants will be able to perform the same function and will be tolerated at the molecular level. However, it is still unknown how these sequence variants spread and become fixed within a population and what the knock-on effects on other sequences may be.

1.5 Consequences of cis-regulatory change

When a regulatory sequence mutates the chance of survival of the new variant module depends on the effect of this change on the function of the module. If a change is deleterious, such as the deletion of a binding site and the gene output is altered so that the organism is compromised then the change will be eliminated from the population. Other sequence changes will produce novel expression patterns and these can be positively selected. If a change is neutral then its fate will be determined by genetic drift and it will spread to fixation or be eliminated and the time this takes to occur will depend on the

population size (Kimura, 1983). If the variant becomes fixed in the population then the sequence would have changed although the function remains the same.

The rapid evolution of regulatory regions combined with the necessity of maintaining the correct output makes it hard to believe that all the changes that occur between closely related species are neutral. For example, the *eve* stripe 2 module of *D. pseudoobscura* differs from *D. melanogaster* in that Bcd binding site 3 has been deleted (see fig. 1.3B), there is a new Kr site and a few of the other sites show sequence changes. Yet the *D. pseudoobscura eve* stripe 2 module drives a conserved pattern of expression in a *D. melanogaster* transgenic (Ludwig *et al.*, 1998). It seems hard to evoke purely neutral changes in the evolution of this interaction and suggests a more complex mode of evolution of these regions.

If a nearly neutral change occurs it could be accommodated by the redundancy of binding sites present within a module. Redundancy could allow for one or more changes to arise within the module but subsequently only compensatory mutations would be tolerated and be selected for. For example, after a mutation that leads to a decrease in affinity of an activating site, there could be selection for an increase in affinity of another activating site or decrease in affinity of a repressing site to maintain the same output from the module. There is evidence for this kind of compensatory *cis*-regulatory evolution from experiments with chimaeric *eve* stripe 2 modules. The proximal half of *D. melanogaster eve* stripe 2 was combined with the distal half of the *D. pseudoobscura eve* stripe 2 module and vice versa (see fig. 1.3C; Ludwig *et al.*, 2000). The expression patterns of these two transgenes showed that although the wild-type modules drive the same expression pattern the chimaeric modules drove expanded and less well defined expression patterns. An explanation for the expanded posterior border within one chimaera could be the absence of the new Kr site along with the reduced potential of an extant Kr site in the *D. pseudoobscura* half of the chimaeric module (see fig. 1.3C; Ludwig *et al.*, 2000).

Such a model of redundancy and compensatory change can explain how the sequences are diverging whilst conserving function between species. The rapidity of the change within regulatory sequences may also be a result of genomic turnover events such as slippage, which are thought to occur more frequently than point mutations (Schug *et al.*, 1998). These genomic processes could account for the rearrangement of conserved binding sites within a module. Furthermore, it is possible that a binding site variant could be spread rapidly through the population by such mechanisms (Dover, 1982). The binding site variant may be slightly disadvantageous but be tolerated because of the redundancy present in regulatory modules. If the variant reached high enough proportions within the population there could be selection in *trans* for an allele of the associated transcription factor, which is more suited to the novel binding site. This selection of a compensatory change in *trans* to maintain the interaction is known as molecular co-evolution (Dover and Flavell, 1984; Ohta and Dover, 1984; Dover 2000). Subsequently, an interaction can diverge between two species because of co-evolution of the interaction within each species (Skaer and Simpson, 2000; Shaw *et al.*, 2002; Ruvinsky and Ruvkun, 2003). For example, hybrids of *D. melanogaster* and *D. simulans* display a variable loss of bristles on the notum. Development of these bristles relies on the correct expression of genes of the achaete-scute complex. The loss of these bristles in hybrids is thought to be as a result of incompatibilities which have arisen between the *trans*-acting factors and *cis*-regulatory elements of the complex between the species (Skaer and Simpson, 2000).

1.6 Studying the evolution of an interaction

Promoter regions are subject to mutational and turnover mechanisms and likely experience the modes of selection discussed above. Discovery of the relative frequency of such events and the contribution of selection to the evolution of an interaction between a transcription factor and target promoter is important. It would involve the comparison of an interaction at the molecular

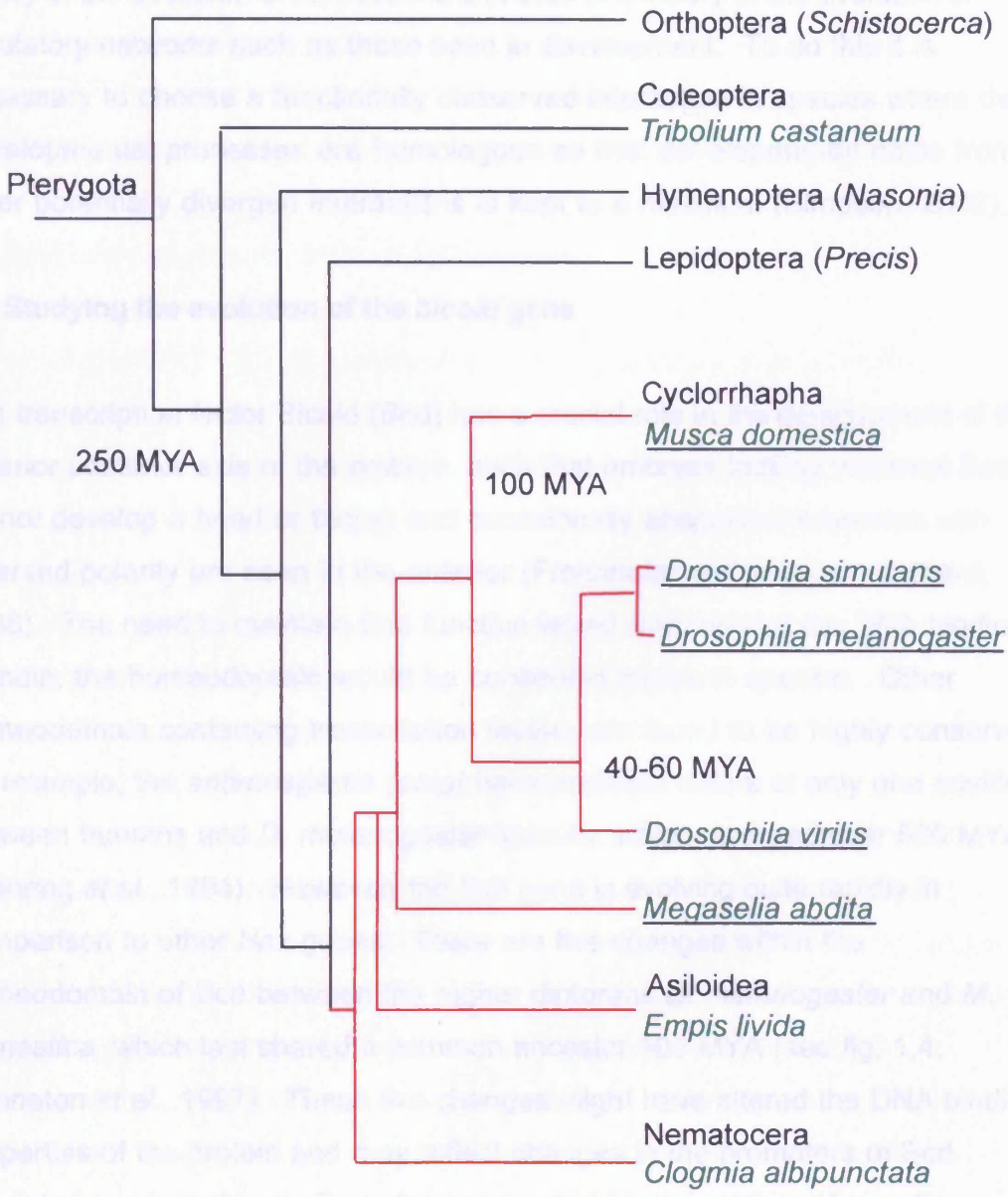


Figure 1.4 Insect phylogeny and relationships within the Diptera
The tree shows the relationship of the Diptera with other insect groups (black lines) and relationships within the Diptera (indicated by the red lines).
Dipteran species from which *bcd* genes have been isolated are underlined.
Individual species referred to in the thesis are shown in green font.
The divergence times are taken from estimates by Beverley and Wilson 1984, and in Richards and Davies 1977. The tree is not to scale and not all branches are shown.

level between two or more species. These studies should help establish a theory of the evolution of interactions and thus to a theory of the evolution of regulatory networks such as those seen in development. To do this it is necessary to choose a functionally conserved interaction in species where the developmental processes are homologous so that developmental noise from other potentially diverged interactions is kept to a minimum (Simpson, 2002).

1.7 Studying the evolution of the *bicoid* gene

The transcription factor Bicoid (Bcd) has a crucial role in the development of the anterior posterior axis of the embryo, such that embryos lacking maternal Bcd do not develop a head or thorax and occasionally abdominal segments with inversed polarity are seen in the anterior (Frohnhofer and Nusslein-Volhard, 1986). The need to maintain Bcd function would suggest that the DNA binding domain, the homeodomain would be conserved between species. Other homeodomain containing transcription factors are found to be highly conserved for example, the *antennapedia (antp)* homeodomain differs at only one position between humans and *D. melanogaster* species, which diverged over 500 MYA (Gehring *et al.*, 1994). However, the *bcd* gene is evolving quite rapidly in comparison to other *Hox* genes. There are five changes within the homeodomain of Bcd between the higher dipterans *D. melanogaster* and *M. domestica,* which last shared a common ancestor 100 MYA (see fig. 1.4; Bonneton *et al.*, 1997). These five changes might have altered the DNA binding properties of the protein and may reflect changes in the promoters of Bcd regulated genes between *D. melanogaster* and *M. domestica.* Hence the changes are an indication that Bcd interactions have evolved between the two species.

A comparison of the expression of a number of genes was carried out between *D. melanogaster* and *M. domestica* (Sommer and Tautz, 1991). This included genes of the maternal, gap, pair-rule, segment polarity and homeotic classes. All genes compared had a conserved expression pattern, the only

differences being in timing of expression or patterns of expression in the later stages of development (Sommer and Tautz, 1991). Therefore, as *D. melanogaster* and *M. domestica* have a conserved mode of early development these species are suitable for a comparison of *bcd* interactions and their evolution (Bonneton *et al.*, 1997; McGregor *et al.*, 2001; Shaw *et al.*, 2001;2002).

1.8 Bcd belongs to a network of interactions

The evolution of the Bcd/*hunchback* promoter interaction was compared between *D. melanogaster* and *M. domestica* and some differences were observed at both the molecular level and in the function of the interaction (see 1.11). Our understanding of the evolution of an interaction such as that between Bcd and the *hb* promoter must be considered within the larger framework of the entire Bcd-dependent gene network (see fig. 1.1A). Bcd regulates many different genes and changes to any Bcd target promoter, whether through drift, genomic turnover or selection could affect its interaction with Bcd and in some circumstances result in a selective change to Bcd or vice versa. Most Bcd regulated promoters are targets of other transcription factors and co-factors and these are also dynamic interactions, which could have an impact on the evolution of Bcd regulation (Small *et al.*, 1991; Liaw and Lengyel, 1992; Hoch *et al.*, 1991). Thus, the general aim of this thesis was to expand the study of the evolution of the Bcd network to include two other Bcd interactions, namely those with the *tailless* promoter and the *caudal* mRNA. This would increase both our knowledge of the *bcd* network at a molecular level and our understanding of the evolution of the Bcd protein and its target promoters between the species of *D. melanogaster* and *M. domestica*.

1.9 The bicoid gene

The *bcd* gene encodes a homeodomain containing transcription factor that acts as a morphogen by regulating the expression of a number of genes along

the anterior posterior axis of the *Drosophila* embryo (see fig. 1.5; Berieth *et al.*, 1988; Driever and Nusslein-Volhard, 1988a,b). *bcd* mRNA is transcribed maternally and deposited in the embryo where it becomes anchored at the anterior tip of the embryo. This requires an element in the *bcd* mRNA 3'UTR and the products of the genes *exuperantia, swallow* and *staufen* (Driever and Nusslein-Volhard, 1988b, St Johnston and Nusslein-Volhard, 1992; Macdonald *et al.*, 1993; Ferrandon *et al.*, 1994). Bcd protein then diffuses posteriorly to set up a concentration gradient of Bcd up to 30% egg length (EL; see fig. 1.5; Driever and Nusslein-Volhard, 1988a,b).

The ability of Bcd to bind DNA and activate transcription was demonstrated on the *hunchback (hb)* P2 promoter (Driever and Nusslein-Volhard, 1989; Struhl *et al.*, 1989). Activation from this promoter results in the expression of hunchback in the anterior half of the egg during the syncytial blastoderm stage of embryonic development (Tautz, 1988). There are seven Bcd binding sites present in the *hb* P2 promoter, with a consensus sequence of TC**TAAT**CCC (the core is in bold, Driever and Nusslein-Volhard, 1989; Driever *et al.*, 1989a).

All *Hox* genes including Bcd contain a DNA binding domain called the homeodomain, which consists of three alpha helices. The third helix known as the sequence recognition helix makes direct contacts with residues in the major groove of the DNA (see fig. 1.6). For example the asparagine at position 51 (asn51), binds to the second adenine (A_3) of the $T_1A_2A_3T_4$ core (Tucker-Kellogg *et al.*, 1997). The residue at position 54 of the homeodomain determines much of the binding site sequence specificity and in most homeodomain containing proteins is a glutamine (see fig. 1.6; Hanes and Brent, 1989, 1991; Hanes *et al.*, 1994; Ades and Sauer, 1995; Tucker-Kellogg *et al.*, 1997). Bcd has a lysine at position 50 and this determines the preference of the Bcd protein for the sequence TAATCC since Lys50 makes direct contact with the cytosine (C_5) of the binding site sequence (Tucker-Kellogg *et al.*, 1997). The bond between Bcd lys50 and C_5 is particularly strong

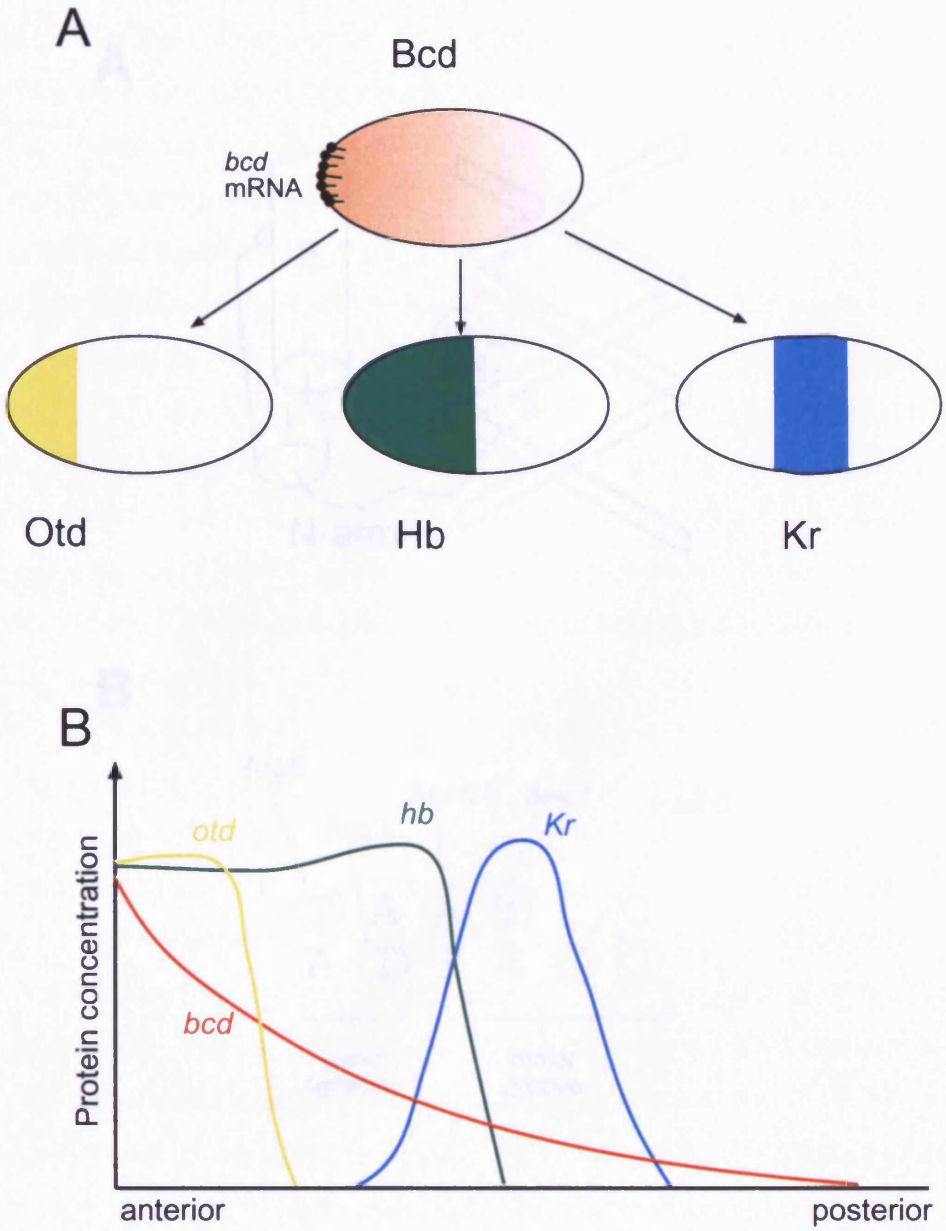


Figure 1.5 The role of *bcd* as a morphogen in early development.
A. bcd mRNA is anchored at the anterior tip of the egg (black circles). Bcd protein then diffuses posteriorly to set up a concentration gradient of Bcd up to 30% egg length (red shading). Bcd activates the expression of a number of genes, three examples are shown: *otd* (yellow), *hb* (green) and *Kr* (blue). Embryos are orientated with the anterior to the left and dorsal up.
B. Bcd activates genes at different concentrations along the anterior posterior axis of the *Drosophila* embryo. The graph shows the concentration of Bcd (y-axis) in relation to the position along the anterior to posterior (x) axis of the embryo. Also shown are the protein concentrations of Otd, Hb and Kr which result from activation by Bcd.

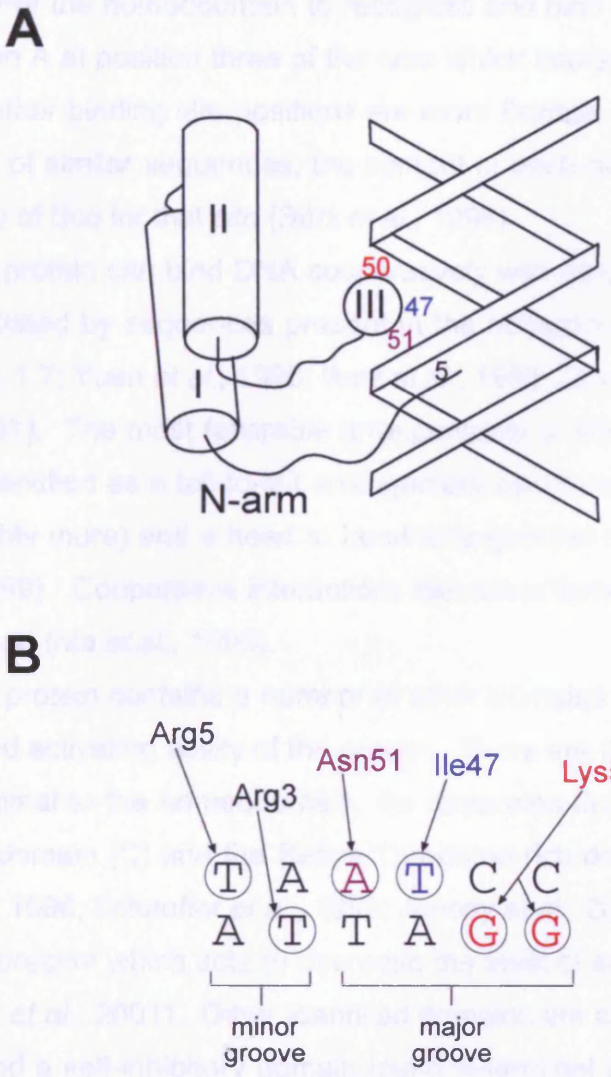


Figure 1.6 Contacts between the Bcd homeodomain and the Bcd binding site sequence
A. The physical relationship between a homeodomain protein and a DNA helix is shown. Homeodomains form three helices (I-III). The third helix lies perpendicular to the other two and makes contact with the major groove of the DNA helix; residues in the third helix are shown in colour. Residues in the N-terminal arm of the protein make contact with the minor groove.
B. The contacts between the residues in the Bcd homeodomain and the Bcd consensus binding site sequence are shown (Tucker-Kellogg *et al.*, 1997). The residues are numbered according to their position within the homeodomain and the arrows indicate which bases they contact and residues of helix three are shown in colour (adapted from Ades and Sauer, 1995, fig. 2).

in comparison to other homeodomain-binding site interactions (Ades and Sauer, 1994). For the homeodomain to recognize and bind a DNA sequence there must be an A at position three of the core which interacts with residue asn51 but the other binding site positions are more flexible. Whilst Bcd is able to bind a range of similar sequences, the content of each binding site does affect the affinity of Bcd for that site (Burz *et al.*, 1998).

The Bcd protein can bind DNA cooperatively with other Bcd molecules and this is mediated by sequences present in the homeodomain and flanking regions (see fig. 1.7; Yuan *et al.*, 1996; Burz *et al.*, 1998; Zhao *et al.*, 2000; Burz and Hanes, 2001). The most favorable arrangements of sites for cooperative binding were identified as a tail-to-tail arrangement separated by 7 to 15 bp (although possibly more) and a head to head arrangement separated by 3 bp (Yuan *et al.*, 1999). Cooperative interactions can occur between sites spaced up to 100 bp apart (Ma *et al.*, 1996).

The Bcd protein contains a number of other domains that influence the DNA binding and activating ability of the protein. There are three activating domains C-terminal to the homeodomain, the glutamine-rich domain (Q), C-terminal acidic domain (C) and the Serine/Threonine-rich domain (ST) (see fig. 1.7; Yuan *et al.*, 1996, Schaeffer *et al.*, 1999; Janody *et al.*, 2001). An A-rich domain is also present which acts to decrease the level of activation of the Bcd protein (Janody *et al.*, 2001). Other identified domains are a PEST domain, a PRD domain and a self-inhibitory domain found N-terminal to the homeodomain, which decreases the activity of the Bcd protein by a factor of 40 (Zhao *et al.*, 2002). This inhibition does not require any other part of the Bcd protein so is thought to prevent interaction between Bcd and either another Bcd molecule or co-activator (Zhao *et al.*, 2002). Indeed Bcd has been shown to interact with Chip a co-activator protein that facilitates enhancer-basal promoter interactions (Torigoi *et al.*, 2000).

Figure 1.7 An alignment of Bcd proteins within the Diptera
The sequences aligned with ClustalW are D. MEL - *D. melanogaster* (Berleth *et al.*, 1998), D. PSE - *D. pseudoobscura* (Seeger and Kaufman, 1990), M.DOM - *M. domestica* (Shaw *et al.*, 2001), C.VIC -*C. vicina* (Mcgregor, 2002) and M.ABD - *M. abdita* (Stauber *et al.*, 1999). The homeodomain is boxed in red, the RNA binding motif in green and the PEST domain in blue. MAP kinase target sites present in the PEST domain and elsewhere are indicated by asterisks (Janody *et al.*, 2000). The black dotted box indicates the PRD domain and the purple box the self-inhibitory domain (see text, Zhao *et al.*, 2002). The black box indicates a serine rich domain seen only in the Calyptratae, putative serine phosphorylation sites are highlighted. The Q-rich and A-rich domains are shown in bold (Schaeffer *et al.*, 1999). In *D. melanogaster* the eIF4e recognition motif is shaded in yellow. The S/T domain is found between the homeodomain and the Q-rich domain and includes the PEST domain (Janody *et al.*, 2001). The C-terminal domain contains most sequences 3' to the A-rich domain (Janody *et al.*, 2001). The boundaries of the S/T and C-terminal domains are not as clearly defined as the other domains in *D. melanogaster* Bcd.

1.10 Functional analysis of the Bcd protein

Functional analysis of *D. melanogaster* Bcd protein indicates it has a weak binding and activating ability in comparison to other *D. melanogaster*, yeast and bacterial transcription factors (Driever *et al.*, 1989b; Ma *et al.*, 1999; Small *et al.*, 1991 and 1992; Arnosti *et al.*, 1996). *in vitro* assays show that whilst Bcd can bind DNA as a monomer activation by Bcd requires more than one Bcd binding site (Burz *et al.*, 1998). It was proposed that the combination of both weak binding ability and activation results in a highly sensitive response to small shifts in the concentration of Bcd or a direct antagonist of Bcd (Small *et al.*, 1991) and this highlights the importance of the cooperative interactions between Bcd molecules. The Bcd binding site sequence is short and appears frequently throughout the genome but these experiments show that unless there are two Bcd sites present in a regulatory module then activation cannot occur. This explains why Bcd does not activate all genes that have a putative Bcd site in their regulatory region (Burz *et al.*, 1998). A result of cooperativity between Bcd monomers is the increase in affinity of Bcd for DNA and this creates a sharp on/off switch for the activation of target genes. Often a high affinity site is found close to a low affinity site and it has been shown that the former site can promote binding at the latter site (Ackers *et al.*, 1983). There is also evidence for a synergistic interaction between Bcd and Hb protein to enhance activation of Bcd targets in the anterior (Simpson-Brose *et al.*, 1994). This synergy involves domains of both the Bcd and Hb proteins, which make contacts with $TAF_{II}110$ and $TAF_{II}60$ of the transcription complex respectively (Sauer *et al.*, 1995a,b, 1996).

The many experiments conducted on the cooperative interaction between Bcd molecules and the variable site recognition and affinity reveal the complexity of the Bcd regulatory activity. This coupled with the gradient of Bcd protein reveals why Bcd is able to participate in the regulation of so many genes throughout the anterior two thirds of the embryo.

1.11 Comparison of the Bcd/*hb* promoter interaction between *D. melanogaster* and *M. domestica*

In both *D. melanogaster* and *M. domestica* Bcd activates expression of *hb* in the anterior half of the embryo (Tautz, 1988; Driever and Nusslein-Volhard, 1989; Bonneton *et al.*, 1997). In *D. melanogaster* this expression is regulated by the P2 promoter, which contains seven Bcd binding sites (see fig. 1.8; Driever and Nusslein-Volhard, 1989). The regulation of *hb* was compared between these two species to look at the evolution of Bcd-promoter interactions (Bonneton *et al.*, 1997). The sequence of the P2 *cis*-regulatory module of *M. domestica* was shown to be unalignable with that of *D. melanogaster* (Bonneton et al., 1997). Bcd binding sites in the *M. domestica hb* promoter were identified by footprinting the region upstream from the transcription start site with the *M. domestica* homeodomain (Bonneton *et al.*, 1997). Ten Bcd binding sites were discovered with a consensus sequence of YTAATCC and the position, sequence, spacing and orientation of these sites is different to those of *D. melanogaster* (see fig. 1.8; Bonneton *et al.*, 1997).

In general it was found that there were more Bcd binding sites in the *M. domestica* promoter and these were spread over a larger region of the DNA (Bonneton et al., 1997). A *M. domestica hb* gene was able to partially rescue a *D. melanogaster hb* mutant. The rescue was less efficient than that seen with a *D. melanogaster hb* construct (Bonneton *et al.*, 1997). This indicates a degree of incompatibility between the *D. melanogaster* Bcd and *M. domestica hb* promoter. Similarly a *M. domestica bcd* gene was able to rescue embryos of *D. melanogaster bcd* mutant mothers to full viability, although the proportion of embryos which survived to adulthood were low in comparison to a *D. melanogaster bcd* transgene (Shaw *et al.*, 2002). A cytoplasm transfer experiment which involved injection of *M. domestica* and *D. melanogaster* anterior cytoplasm into the anterior of *D. melanogaster bcd* mutant embryos resulted in a higher degree of survival of embryos injected with the latter (Schroder and Sander, 1993). These experiments are in agreement with a

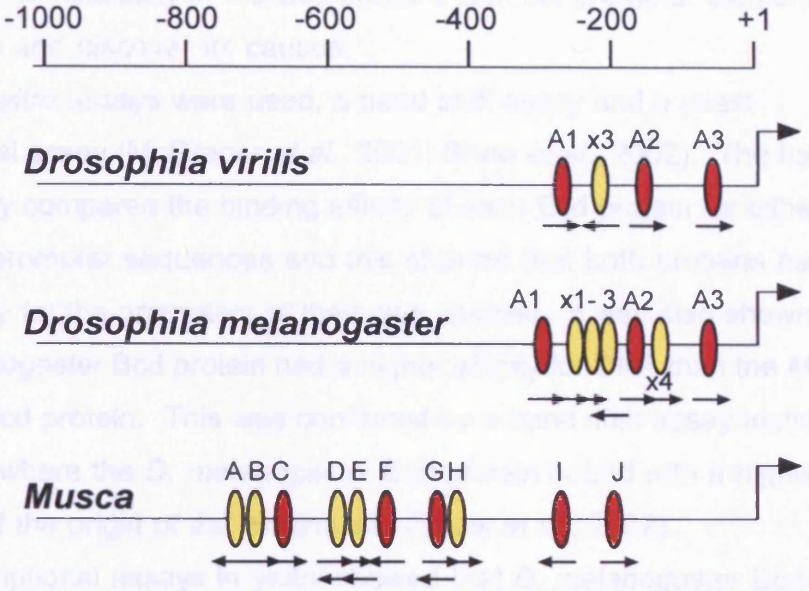


Figure 1.8 Comparison of Bcd-dependent *hb* promoters in higher Dipterans. The cartoon shows the different structures of the *hb* P2 promoter in *D. melanogaster*, *D. virilis* and *M. domestica* species. The large arrow is the transcription start site. The numbered bar represents the distance in bp 5′ from the transcription start site. Red ovals represent Bcd-binding sites with a canonical core sequence (TAAT), while the yellow ovals represent sites with a non-canonical core sequence (TAAG, AAAT, CAAT, TCAT and TGAT). Smaller arrows represent the orientation of sites. All binding sites are labeled according to previously published *DNaseI* footprinting data (*D. melanogaster* - Driever and Nusslein-Volhard, 1989; *D. virilis* - Lukowitz *et al.*, 1994 and *M. domestica* - Bonneton *et al.*, 1997).

level of incompatibility in the Bcd-*hb* promoter interaction between the two species. Therefore, a number of functional studies were carried out to assess the level of incompatibility of the Bcd proteins with the promoter elements of both species and discover its causes.

Two *in vitro* assays were used, a band shift assay and a yeast transcriptional assay (McGregor *et al.*, 2001; Shaw *et al.*, 2002). The band shift assay directly compared the binding affinity of each Bcd protein for either species *hb* promoter sequences and this showed that both proteins had a higher affinity for the promoters of their own species. It was also shown that the *D. melanogaster* Bcd protein had a higher affinity for DNA than the *M. domestica* Bcd protein. This was confirmed by a band shift assay testing single sites, where the *D. melanogaster* Bcd protein bound with a higher affinity regardless of the origin of the binding site (Shaw *et al.*, 2002).

Transcriptional assays in yeast showed that *D. melanogaster* Bcd activated transcription from the *D. melanogaster hb* promoter to a higher level than the *M. domestica* Bcd protein. This result could be expected if there are incompatibilities in the interaction between the two species. In addition, levels of transcription from the *M. domestica hb* promoter were similar for both Bcd proteins. This could be a result of the strong affinity of *D. melanogaster* Bcd for DNA, which could overcome the incompatibilities that have evolved between the species. Importantly, the *M. domestica* Bcd protein was able to activate an equivalent level of transcription from the *M. domestica* promoter as the *D. melanogaster* Bcd protein but activated the *D. melanogaster* promoter more weakly than *D. melanogaster* Bcd. The overall interpretation of the *in vitro* assays was that the Bcd proteins of both species could bind and activate expression from the other species promoter but not to the same level (Shaw *et al.*, 2002).

Both the *in vivo* and *in vitro* experiments show that despite the differences between the Bcd proteins and the *hb* promoter regions the interaction is still functional between the two species. However, it was apparent that the mixed species interactions such as between *D. melanogaster* Bcd and the *M.*

domestica hb promoter were not equivalent to the wild type interaction within each species. Therefore, there may have been co-evolution of the interaction within each species since their divergence.

1.12 Bcd regulates *tailless* expression

Another target of Bcd *tailless (tll)* encodes a transcription factor that is essential for the specification of the terminal regions of the embryo (Pignoni *et al.,* 1990, 1992). Bcd regulates the expression of *tll* in an anterior stripe at the syncytial blastoderm stage of development (Liaw and Lengyel, 1992). The terminal system and Dorsal also regulate *tll* and thus the regulatory regions of *tll* are more complex than those of *hb* (Liaw and Lengyel, 1992). The *tll* regulatory region has been partially characterized in *D. melanogaster* and consists of at least 5 kb of sequence 5' to the transcription start site of the *tll* gene (Liaw and Lengyel, 1992). The modules that drive the different domains of expression of *tll* are overlapping and each contain multiple binding sites for a number of regulatory factors (see fig. 1.9). Within the *tll* promoter sequence regions responsive to Bcd regulation have been identified (see fig. 1.9). The regulation of *tll* by Bcd is further complicated because of a regulatory interaction between Bcd and the terminal system, which is not yet fully understood (Ronchi *et al.*, 1993; Janody *et al.*, 2000, 2001).

The terminal system is a signal transduction cascade that directs patterning in the non-segmented head and tail regions of the embryo (Nusslein-Volhard, 1987; for review see Perrimon, 1993). The action of the terminal system is mediated through the Tor RTK protein (Tor) which when activated leads to the derepression of *tll* (Liaw *et al.*, 1995). Unfortunately, the direct regulator of *tll* has remained elusive although sequence motifs responsive to Tor regulation, the Torso response element (Tor-RE), have been identified within the *tll* promoter (Liaw *et al.,* 1993, 1995). A number of factors bind to the Tor-RE including NTF-1, Capicua (Cic), and GAGA (Liaw *et al.,* 1995; Jimenez *et al.,* 2000; Chen *et al.,* 2002). Interestingly, Cic is able to bind the co-

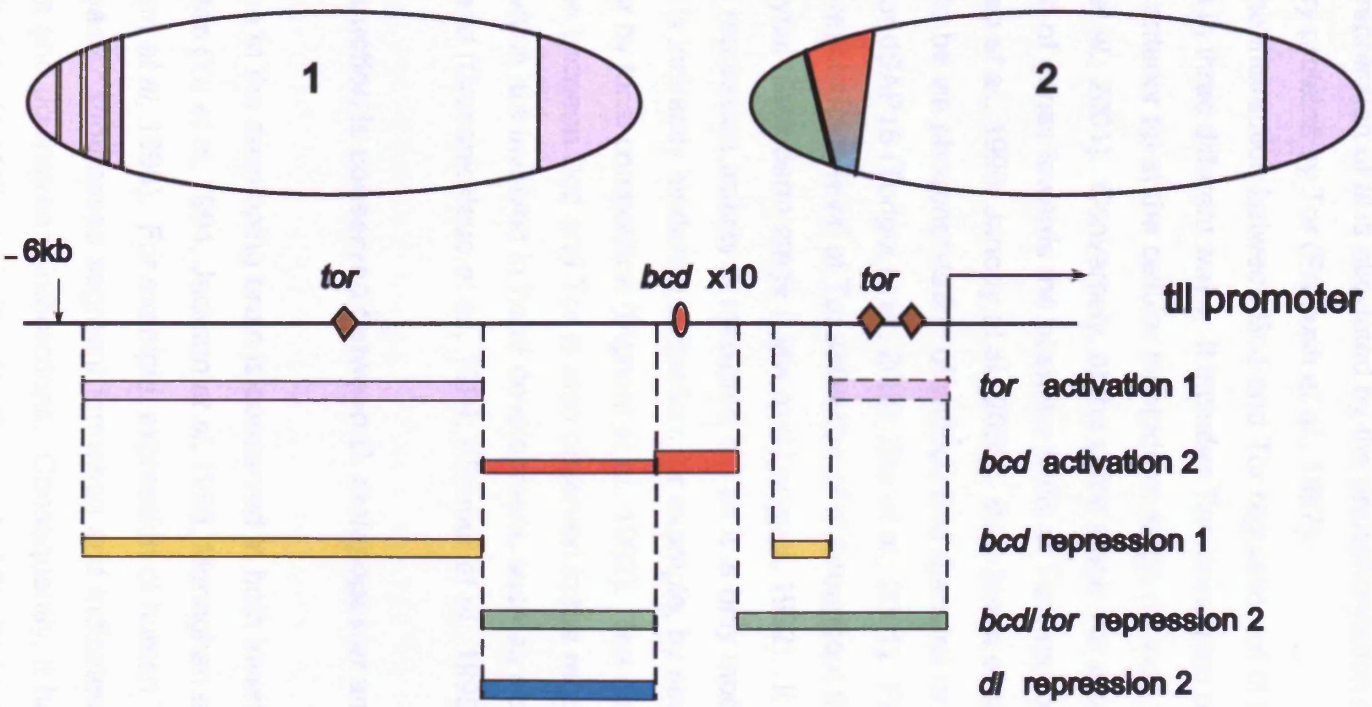


Figure 1.9 The expression patterns of *tll* in *D. melanogaster* and the *tll* promoter structure
A. *tll* expression patterns in the anterior and posterior of syncytial and cellular blastoderm embryos. Embryos are orientated with the anterior to the left and dorsal up. The coloured areas represent expression domains regulated by Tor (purple), Bcd (red and yellow), Dl (blue) and Tor and Bcd (green).
B. The structure of the *tll* promoter, the large arrow represents the transcription start site and the length of the sequence is shown in kb. Footprinted binding sites for Tor (diamonds) and Bcd (oval) are shown. The *tll* promoter contains regions that are responsive to Bcd, Tor and Dl regulation. These regions are overlapping and are colour coded to show the domains of *tll* expression that result and are shown in **A**. Numbering refers to the stage of development at which the regulation occurs:1. syncytial and 2. cellular blastoderm.

repressor Groucho (Gro) *in vitro* and it is thought that Cic recruits Gro to the *tll* promoter to repress activation of *tll* (Paroush *et al.*, 1997). There is evidence that the repression of *tll* is alleviated by the phosphorylation of one of these regulatory proteins by Tor (Paroush *et al.*, 1997).

The interaction between Bcd and Tor regulation of *tll* is complicated and operates in three different ways. It includes Tor repression of Bcd activation of *tll* at the anterior tip at the cellular blastoderm stage (Ronchi *et al.*, 1993; Janody *et al.*, 2001). Conversely, at the same stage Tor also enhances Bcd activation of genes towards the posterior limits of Tor function (Bellaiche *et al.*, 1996; Gao *et al.*, 1996; Janody *et al.*, 2000). Both forms of regulation are thought to be via phosphorylation of various Bcd domains or of a co-factor such as Chip or dSAP18 (Torigoi *et al.*, 2000; Zhu *et al.*, 2001). Finally, Bcd appears to down-regulate the level of Tor activation of *tll* throughout the anterior during the syncytial blastoderm stage (Liaw and Lengyel, 1992). It is not understood how this repression activity is mediated but as it is only moderate it suggests that Bcd is indirectly hindering activation: for example, by sequestering co-factors or by local competition (Pignoni *et al.*, 1992). The complex cross regulation between Bcd and Tor is also observed in the regulation of other genes, which are involved in head development, such as *sloppy-paired1/2* and *buttonhead* (Grossnicklaus *et al.*, 1994; Wimmer *et al.*, 1995).

1.13 *tll* function is conserved between *D. melanogaster* and *M. domestica*

tll function in the developing brain is conserved in both invertebrates and vertebrates (Yu *et al*, 1994, Jackson *et al*, 1998, Monaghan *et al*, 1995, Hollemann *et al*, 1998). For example, expression of human Tlx in *D. melanogaster* suppresses segment formation and indicates conservation of upstream and downstream interactions. Consequently, it has been proposed that the initial role of *tll* was in the development of the brain and was later recruited to the posterior presumably because of its regulation by the terminal system (Rudolph *et al.*, 1997). Therefore, it is likely that the regulation by Bcd

and the other regulatory systems are conserved between *D. melanogaster* and *M. domestica*. Indeed a cursory study of the expression of *tll* in *M. domestica* revealed the expression pattern seen in the cellular blastoderm stage to be identical to that seen in *D. melanogaster* (Sommer and Tautz, 1991).

1.14 Dipterans and emergence of Bcd

bcd and its sister gene *zen* are the result of a duplication of an ancestral *Hox3* gene (Stauber *et al.*, 1999). This duplication occurred within the dipteran lineage before the radiation of the cyclorrhaphan flies (see fig. 1.4; Stauber *et al.*, 2000; Brown *et al.*, 2001). The *Hox3* gene sequences of dipterans basal to the cyclorrhapha most closely resemble *zen* but are expressed maternally and zygotically in domains reminiscent of both *bcd* and *zen* in *D. melanogaster* (Berleth *et al.*, 1988; Rushlow *et al.*, 1987; Stauber *et al.*, 1999, 2002). This suggests that *bcd* and *zen* were originally expressed both maternally and zygotically. *zen* is involved in the specification of the extra-embryonic membranes in *D. melanogaster* and the loss of maternal *zen* expression in the stem lineage of the cyclorrhaphan flies correlates with both a reduction in the domain of *zen* expression and size of the extraembryonic tissue in *D. melanogaster* (Rushlow *et al.*, 1987; Stauber *et al.*, 2002). The reduction of extraembryonic tissue has resulted in more of the egg being dedicated to the developing embryo and could have facilitated the evolution of the long germ band mode of embryogenesis. In this mode of development the embryo develops along the entire anterior-posterior axis of the egg and all segments are specified by gastrulation (Sander, 1975). In species that exhibit the ancestral short germ band mode of development the anterior of the egg forms extraembryonic tissue whilst the embryo develops in the posterior. Thus, in comparison to long germ band insects the position of the segments within the egg and the timescale in which they develop is different (for review see Nagy, 1994). This change in development between long and short germ band insects complicates comparative studies of these species, which are

necessary to determine the origins of *bcd* (Dearden and Akam, 1999).

It has recently been shown that in *T. castaneum,* a short germ band insect that the genes *hb* and *otd* specify patterning of the head and thorax (see fig. 1.4; Schröder, 2003). There is direct evidence to suggest that *bcd* became the anterior determinant by assuming the regulation of *hb* and *otd*. Firstly, *bcd* regulates the expression of both *otd* and *hb* (Gao and Finkelstein, 1998; Driever and Nusslein-Volhard, 1989). Secondly, Bcd and Hb share the regulation of a number of target genes, although, in *D. melanogaster* the role of Hb is seen as an accessory to Bcd (Simpson-Brose *et al.*, 1994; Arnosti *et al.*, 1996). Thirdly, during the evolution of the Bcd protein the homeodomain has assumed the same recognition helix as the Otd protein (Hanes and Brent, 1989; Finkelstein *et al.*, 1990). Therefore, Bcd would have become able to bind Otd binding sites and thus regulate Otd target genes. As *bcd* is expressed in the anterior of embryos this may have been one of the key innovations that enabled the long germ band mode of embryogenesis to evolve in the cyclorrhapha (Stauber *et al.*, 2002). Importantly, the evolution of these new functions of Bcd may explain the rapid evolution of the *bcd* gene sequences in comparison to *zen* (Stauber *et al.*, 1999; Baines *et al.*, 2002).

1.15 Bcd regulates translation of *cad* mRNA

The Bcd homeodomain is unusual in that it is able to bind RNA as well as DNA and this function requires lys50 of helix three (Dubnau and Struhl, 1996; Rivera-Pomar *et al.*, 1996). The third helix also contains a motif that resembles the ARM domain of the RNA binding protein Rev of HIV-1 (Heaphy *et al.,* 1990). Deletion of arg54 within this motif destroys the RNA binding ability but not DNA binding ability of Bcd (Niessing *et al.,* 2000). Bcd binds the mRNA of *caudal (cad)* in the syncytial blastoderm and represses its translation (Dubnau and Struhl, 1996; Rivera-Pomar *et al.*, 1996). The *cad* mRNA is homogeneously distributed throughout the egg but Bcd repression in the anterior results in a gradient of Cad protein which is highest at the posterior (Mlodzik *et al.,* 1985;

Macdonald and Struhl, 1986). The repression of *cad* mRNA translation is necessary because the presence of Cad in the anterior of embryos causes ectopic head structures to develop (Niessing *et al.,* 1999).

Bcd binds a 120 nt element in the 3'UTR of the *cad* mRNA called the Bcd responsive element (BRE; Dubnau and Struhl, 1996; Rivera-Pomar *et al.*, 1996). Repression of translation requires the PEST domain of Bcd and deletion of key serine and threonine residues results in loss of repression (Niessing *et al.,* 1999). Another motif just N-terminal to the homeodomain is also required for repression, this enables Bcd to bind directly to the translation initiation factor 4E (eIF4E; Niessing *et al.,* 2002). The eIF4E protein binds the 5' cap of mRNA and this interaction is required for translation initiation (Sachs and Varani, 2000). A Bcd mutant protein lacking the eIF4E binding motif is unable to repress translation of *cad* mRNA but is still able to activate transcription of *hb* (Niessing *et al.,* 2002). It is likely that Bcd binds the eIF4E protein and prevents the interaction of eIF4E with other translation initiation factors either by direct competition or by eliciting a structural change in the eIF4E protein (Niessing *et al.,* 2002). Another protein, Bin3, which has been shown to bind to Bcd shows similarity to protein methyltransferases and contains a SAM binding motif which enables methylation of DNA and RNA (Zhu and Hanes, 2000). This protein may be involved in the regulation of *cad* mRNA by Bcd.

Neither *Hox3* genes nor *zen* contain an RNA binding motif therefore the Bcd RNA binding ability has evolved since the duplication of the *Hox3* gene in the dipteran lineage (Stauber *et al.*, 1999; Rivera-Pomar and Jackle, 1996). Bcd is unable to regulate *cad* mRNA of the lower dipteran *Clogmia albipunctata*, yet in *T. castaneum* and the silk worm, *Bombyx mori*, a gradient of Cad protein is seen albeit later in development than in *D. melanogaster* (see fig. 1.4; Rivera -Pomar *et al.*, 1996; Xu *et al.*, 1994; Schulz *et al.*, 1998). This suggests that translational regulation of *cad* mRNA was present prior to the emergence of the diptera but that the regulatory element was different to that recognized by Bcd (Schröder, 2003). This indicates that *bcd* assumed the role

activation of anterior determining genes and the repression of posterior determining genes.

1.16 The aims of the thesis

The Bcd-*hb* promoter interaction has been compared between *D. melanogaster* and *M. domestica* and it has been shown that the function of the interaction is conserved between the species (Bonneton *et al.*, 1997; Shaw *et al.*, 2001; Shaw *et al.*, 2002). Yet there is divergence in the interaction both in the promoter sequences and the Bcd protein. Functional studies of the Bcd-*hb* promoter interaction have begun to dissect the meaning of these sequence changes between the two species (Bonneton *et al.*, 1997; McGregor *et al.*, 2001; Shaw *et al.*, 2001; Shaw *et al.*, 2002). However, these changes could reflect or influence other Bcd interactions within the developmental regulatory network and so far this possibility has not been examined. Comparison of further *bcd* interactions between the species will determine if the changes observed between species for the Bcd-*hb* interaction are typical of other Bcd-promoter interactions. This approach may also highlight the changes that could be responsible for the evolution of the *bcd* gene. Therefore, the aim of this thesis is to widen the comparison of Bcd interactions between *M. domestica* and *D. melanogaster,* to provide a comparison for the *hb* promoter and begin to understand the evolution of an interaction within a regulatory network.

The major questions that were addressed in this thesis are as follows:

1 Is the *tll* gene conserved in structure and function between the two species?

2 Is there evidence for a Bcd regulatory region within the *M. domestica tll* promoter and how does this compare, structurally and functionally with that of *D. melanogaster*?

3 Is the Bcd/*tll* promoter interaction co-evolving between *D. melanogaster* and *M. domestica*?

4 Are the regulatory regions of *M. domestica tll* recognized by *D. melanogaster*

Bcd *in vivo*?

5 Is it possible to identify the mechanisms by which non-coding sequences evolve and can this explain the differences observed between the regulatory regions of the two species?

6 Is the *cad* gene conserved in structure and function between the two species and is *M. domestica cad* mRNA regulated by Bcd in the same way as *D. melanogaster cad*?

Chapter 2 Materials and Methods

21 Materials

21.1 Media

LB (Luria broth): 1% (w/v) Bacto-tryptone (Difco), 0.5% (w/v) Bacto-yeast extract (Difco), 1% (w/v) NaCl. LB-agar was made as above but with the addition of 1.5% (w/v) agar (Difco). The antibiotics ampicillin and kanamycin were added when appropriate to LB cultures and LB-agar plates to a working concentration of 50 µg/ml (from stock solutions of 50 mg/ml in ethanol). Tetracycline was used at a working concentration of 12.5 µg/ml (stock solution 12.5 mg/ml in 50% aqueous ethanol).

Oat food (per litre): 130 g ground oatmeal, 6 g agar, 40 ml black treacle, 5.5 ml 20% (w/v) Nipagin.

Sugar food (per litre): 46.3 g sucrose, 7.1 g agar, 82.2 g dried yeast, 10 ml 20% (w/v) Nipagin.

21.2 Organisms

Bacteria

The following strains of *Escherichia coli* were used:

DH5α (Gibco BRL) *supE*44 *hsdR*17 *recA*1 *endA*1 *gyrA*96 *thi*-1 *relA*1.

XL1-Blue (Stratagene): *supE*44 *hsdR*17 *recA*1 *endA*1 gyrA46 *thi relA*1 *lac*$^-$ F′ [*proAB*$^+$ *lacI*^{q} *lacZ*ΔM15 Tn*10*(tet^r)].

XL-1 Blue MRA: Δ(*mcrA)183*, Δ(*mcrCB-hsdSMR-mrr*)*173, endA1, supE44, thi-1, gyrA96, relA1, lac.*

Bacterial stocks and stocks transformed with plasmids were maintained at -20°C in equal volumes of overnight LB cultures and glycerol.

D melanogaster

Stocks used were w118, SbeΔ2-3/TM6, w; ScO/CyO; MKRS/TM6, tor4/CyO, Df(3R)LIN/TM6 (DfLIN,st,Pp,e,bcd$^-$) and w;bcdE1/TM3. All stocks of *D.*

melanogaster came from the Bloomington stock centre, Indiana, USA; except w;bcdE1/TM3 which was donated by M. Stauber, Max-Planck-Institut für biophysikalische Chemie, Göttingen. All stocks were maintained on sugar and oat food.

M domestica

Laboratory strains of *Musca* were donated by the following sources: Cardiff and Rentokil; Dr L. Senior, Insect Investigations, University of Cardiff. Millan, Scott, White and Zurich; Prof. A. Dubendorfer, University of Zurich. Rutgers; Prof. Plapp, University of Arizona. Flies were maintained in cages at 26°C with sucrose, dried milk and water. Larval food was prepared as described in Bonneton *et al.*, 1997.

21.3 Plasmids

Plasmid	Description	Source
pbcdTN3	*D melanogaster bcd* residues 85-166	P. Shaw
pBCDR1	*Musca bcd* residues 59-160 with *Eco*RI and *Hin*dIII flanking sites in pBluescript KS+.	P. Shaw
pCaSpeR AUG _gal	P-element vector containing the *white* and *lacZ* genes and a start codon for translation of the reporter enzyme	Thummel *et al.*, 1988

Table 2.1 Plasmids used in this work.

21.4 Oligonucleotides

All oligonucleotides used were synthesised by Interactiva Biotechnologie and supplied as lyophilised pellets. The sequences are listed in table 2.2.

Name	Tm	Sequence 5' to 3'	Use
MTAIL1	56	CAT CAT GGC ATC TTT GTA CAT G	sPCR
MTAIL2	60	TGG ACA GCA TCT TTG TTC ATG C	sPCR, *In situ* / Southern probe and Intra-specific analysis
MTAIL3	60	GCT AAT TCA CCA CCG TCC TCG TCC	sPCR
MTAIL4	55	CGA TCA TCC CCT TTT ACC TAG	sPCR
MTAIL5	56	GCG TGT TGA TGT AAT TAT TGG G	sPCR
MTAIL6	60	GGG ATT AAA GGG AAA GCA ATT GA	sPCR
5RTLL1	42	TGG AAC GCT TAA AGA AA	5'RACE
5RTLL2	62	CCA GCA CAG CCA TCA CAT GCA TA	5'RACE
5RTLL3	59	GCC ATC ACA TGC ATA GAT ACC GTA A	5'RACE
TLLSUB1	60	CTC AAC GGA AAA TAT CTC AAG TAT GAG ATT T	Sequencing
TLLSUB2	63	GAC AAA AAC ACG GCA GAG TGG CAT AAA	Sequencing
TLLSITU1	60	TGG CTG TGT GTG TAA CCA ATG AA	*In situ* / Southern probe
TLLL3	60	TCG CCA GAC AAT AAC CCT GT	Sequencing L
TLLL7	60	GGG TTA CCA AAC CGG TAA CA	Sequencing L
TLLS7	60	GAC CAT CTG GCC ATT GCT AT	Sequencing S
TLLM3	60	GCT TGT GCC GAC TTG GTA A	Sequencing M
TLL87	60	ATG ACA TCG GTC TTC CGA AG	Sequencing
TLL83	60	ATA ATG GCT GCC GAA CAC AT	Sequencing
PST2KA	50	TTC AAG CTG TGC GAA ACG	Sequencing S
PST2KB	50	AAT CGT GCC AAA GTA GAC C	Sequencing S
PSTVP	50	GTC TAC CGT CTT CGT ATA GC	Sequencing S
PSTVE	50	GAC ATA ATC CGC AAA ATC C	Sequencing S
LAMV3	50	TGT TTA CGC TCT CTG TCG	Sequencing M
LAMV7	50	AGT GAC CGA ATG TCA TCG	Sequencing M
LAM23	50	TGA TAG GAG CGG AGA TCG	Sequencing M
LAM27	50	TTA CCT TTA CAA TTA AGC TTG C	Sequencing M
TLLM7C	50	GGT TTA TAC GAA ACA GTC TCA AG	Sequencing S
TLLSA	50	AGG AAA CAA TAA TCC CAA TAG C	Sequencing S
TLLSB	50	AAA GAA GAA TGC TTT CCT GC	Sequencing S

TLLL1	50	TTC TTC ACT GTC CAC CAG C	Sequencing L
MTLL5.1	50	ATT CAA TTT TAT TCC GAA CAT AGG	Sequencing M
MTLL3.1	50	TAG CAA TCA TCG TTA TTA TGC TC	Sequencing M
MTLL5.2	50	TAG TTC TAA TAA GTG TTA TTA ACG G	Sequencing M
MTLL3.3	50	ACA AAT TTA TTG AAA GTT GCT GTC G	Sequencing M
MSTR2F	50	TTA GGT CGC ATC CTA TAT CAT GTG C	Intra-specific analysis
MSTR2R	50	CAC GAA TGA GCT CAT ATT CAT GGC	Intra-specific analysis
MSTR3R	50	CAT GTG ATG GGA TCC TGG	Intra-specific analysis
MSTR5	50	ATG TTT CAC TGG TGT ATC GAC C	Intra-specific analysis
M3R1	60	AGT TCA TGC CTT CCA AGA TGT CC	3'RACE
M3R2	57	ATA GCC ATG AAT ATG AGC TCA TTC	3'RACE / Intra-specific analysis
DNA1	50	TGC ACT CTA CCA TTC ATA CGG	Footprinting
DNA2	50	ATG TAT GCG AAT ATA CAC ACG	Footprinting
DNA3	50	CAT TCG TGT ACG TGT GTG TGG	Footprinting / Intra-specific analysis
DNA4	50	TCA GTT GAT GGA AGA GCA GC	Footprinting
DNA5	45	CTT GGA ATT AAT TGT CGA TGG	Footprinting / Intra-specific analysis
DNA6	45	TCA ATT GCT TTC CCT TTA ATC C	Footprinting / Intra-specific analysis
DNA7	45	AAT TGT ATG AGT CCG CAT G	Footprinting
DNA8	45	TAT GAC ATC GGT CTC CCT AAG	Footprinting / Intra-specific analysis
DNA9	50	TGG AAT TCT TAC AAA ATA TGC	Footprinting
DNA10	50	ATC AAA TAT GGG ATA AAG CCT G	Footprinting / Intra-specific analysis
DNA11	45	AGG CTT TAT CCC ATA TTT GAT AC	Footprinting
DNA12	45	AAT GCC ACA GAA ATG TCC	Footprinting / Intra-specific analysis
BCD1	50	ACA CAT CTT ATA CAT CCA CTT GG	Footprinting
BCD2	50	TAA CAG TGT TGA AAT CTA GGT CC	Footprinting
BCD3	50	TAT GTC GGA ATT TGG AGT AGG	Footprinting

BCD4	50	ATA ACA TGG TCG GCC TGC	Footprinting
BCD5	50	AGT TTA ACT TGC ATT AGC ATG C	Footprinting
BCD6	50	ATA CTT AGT CGT GAC AAG GTA GC	Footprinting
BCD7	50	TAA GGT TGG CAT GCA CTG	Footprinting
BCD8	50	TCA ACA TAA AGC GCT ATT GG	Footprinting
DMTLL1	50	TTG GTT AGCAGA AGT TATTCC	Footprinting
DMTLL2	50	ATC TGA GTA TGA ATT TTG TAT CG	Footprinting
DMTLL3	50	TTA CGA TAC AAA ATT CAT ACT CAG	Footprinting
DMTLL4	50	TTC GAG TGG CGA TAG TAG C	Footprinting
DMTLL5	50	TAG AAG CGA ACC CAC AGG	Footprinting
DMTLL6	50	TTT CCG CAG ATT CAC TAC C	Footprinting
DMTLL7	50	ATT GAG AAT GAG AAT GAG CG	Footprinting
DMTLL8	50	TTG TCT GCT GTG AGG ACC	Footprinting
TLLP1F	58	ACG GGA TCC CGG CGG TGT TGC AGA TTC TTA GTG GAT T	Promoter insert for transgenic
TLLP1R	58	CTG AGG GGT ACC CCG CTC TGT TTG AGT TGT GTT C	Promoter insert for transgenic
LACZF	50	AAC TTA ATC GCC TTG CAG C	*In situ* of transgenic
LACZR2	50	TTC AGA CGTAGT GTG ACG	*In situ* of transgenic
PLAC4	60	ACT GTG CGT TAG GTC CTG TTC ATT GTT	Inverse PCR of transgenic gDNA
PLAC1	60	CAC CCA AGG CTC TGC TCC CAC AAT	Inverse PCR of transgenic gDNA
CADF	45 to 35	ARG AYA ART ACC GCG TRG TRT AC	Degenerate PCR
CADR1	45 to 35	TTR GCR CGR CGR TTY TGG AAC CA	Degenerate PCR
MCAD	63	TAC TGT ACA TCA CGC TAC ATC ACC	sPCR
MCADN	53	ACG CGT CGA CGC TAT CGC TGT CGG AG	sPCR
CADPR1	52	TCC AGC ATC CCG ACT ATG CC	Southern probe
CADPR2	52	TGA GTG CTG CCA TTC ACA TC	Southern probe
CAD5PRI1	56	GGT GAT GTA GCG TGA TGT ACA GTA	sPCR / RT-PCR
CAD5PRI2	60	CAT AGT CGG GAT GCT GGA CAC C	sPCR
CAD3R3	54	AGT ATG TCA ACT GCA AGT TGA A	3'RACE
CAD3R4	58	CAT CGC ACT GCC ACA CAG CTT AGA TA	3'RACE

CAD3R5	52	GGT CAA TAT CCT AAC AGG G	3'RACE
CAD3R6	55	ATC AAA CAT TTG TAG CCG TCC	3'RACE
CAD 3R7	58	GCC CCC TCT TTG TAT TTA TAA GTG AGA AA	3'RACE
CAD3R8	62	GTT CAA TAC TGT GCA ATT ATC TAT AAC TAC AAC ACA	3'RACE
CADSITU1	54	CCC GCA CCA AGG ATA AAT A	*In situ* probe
CADSITU2	54	CCA ATG CAC TTT CAA CAT CAT	*In situ* probe
RTCDF6	60	CCC ATH CAN GTN AGY GGH	RT-PCR
CAD5A	65	TGG CTG TCG ATG GAC TAT GG	sPCR
CAD5B	60	ATG ATG GGC AAT GTG ATT CG	sPCR
CAD5D	55	CTT ACT TCG CCG GAC AAC C	sPCR
AOL995	60+	CGCGTTTTGTCGACGAATTCTTTC	sPCR

Table 2.2 Primers used in this work.

22 Methods

22.1 Standard molecular biology techniques

22.1.1 DNA precipitation and phenol-chloroform extraction

Alcohol precipitation and phenol-chloroform extraction of nucleic acids were done according to Sambrook *et al.*, (1989).

22.1.2 Restriction digests

Restriction digests were done according to the manufacturer's recommendations in the buffer supplied with the enzyme. Vector DNA was dephosphorylated by the addition of 2-3 units of Shrimp Alkaline Phosphatase (SAP) (5 units/µl, USB-Amersham) to the restriction digest.

22.1.3 Gel extraction

Fragments were run out on standard agarose gels in 1x TAE and gel-purified using a gel extraction kit (Qiagen) according to the manufacturers instructions.

22.1.4 Ligation of DNA fragments.

10-50 ng of linearised vector DNA was incubated with an appropriate amount of insert DNA to give a rough molar ratio vector:insert of 1:3. The DNA was then put on ice and T4 ligase buffer (supplied with enzyme) added to 1x concentration (Gibco BRL), with 1-2 units of T4 DNA ligase (1 Weiss unit/µl, Gibco BRL). The reaction was incubated overnight at 16˚C in a final volume of 15 µl. Half of the ligation reaction was transformed into *E. coli* as described below.

PCR products were cloned and transformed into *E. coli* using TOPO kits (Invitrogen) according to the manufacturers instructions.

22.1.5 Transformation of *E. coli*

Electroporation was used for transformation of plasmids. Electrocompetent cells were prepared as follows: a 0.5 l LB-tetracycline culture of *E. coli* strain XL1-blue was grown to mid-log phase (OD_{600}=0.55). Cells were washed sequentially to remove salts as follows: Cells were pelleted by centrifugation in a Sorvall ultracentrifuge GS-3 rotor at 4000 rpm for 10 mins (4°C). The cell pellet was resuspended in 500 ml of ice-cold deionised water. The cell suspension was spun again and the cell pellet resuspended in 250 ml of ice-cold deionised water. The cell suspension was then spun in a SS-34 rotor at 9000 rpm and resuspended in 10 ml of ice-cold 10% (w/v) glycerol. The cell suspension was then spun again and the pellet resuspended in 1 ml of ice-cold 10% (w/v) glycerol. 40 µl aliquots of cell suspension were frozen in dry ice-ethanol. Cell aliquots were stored at -80°C.

Electroporation: plasmid DNA was prepared by ethanol precipitation and resuspended in 10 µl of deionised water. An electrocompetent cell aliquot was thawed on ice and added with the transforming DNA to an electroporation cuvette. An electric pulse was delivered using a slot apparatus unit (GenePulser, Biorad), set at 25 µF and 1.5 kV. Cells were recovered at 37°C for 1 hour in 1 ml of SOC medium (prepared as described in Sambrook *et al.*, 1989). Aliquots of recovered cells were plated out on appropriate agar medium. Typical efficiency: 1 x 10^8 transformants per µg of DNA.

22.1.6 Preparation of plasmid DNA

Plasmid DNA was isolated from bacterial cultures using mini- or maxi-prep kits (Qiagen) according to the manufacturers instructions.

22.1.7 Agarose gel electrophoresis

0.5-1.8% (w/v) gels were cast using Seakem LE agarose (Flowgen) dissolved in 1x TAE. 5x loading buffer (5x TBE, 15% (w/v) Ficoll-400 (Pharmacia Biotech.), 0.25% (w/v) bromophenol blue) was added to the DNA samples before loading and gels were run in horizontal perspex slab gel tanks at 1-6 V/cm in the

corresponding buffer. DNA was visualised by the addition of ethidium bromide (EtBr) (0.5 µg/ml) to the gel mix before casting and observing the fluorescence at 300 nm UV on a transilluminator. DNA size markers, such as λ*Hin*dIII markers (Gibco BRL) (fragments 23130, 9416, 6557, 4361, 2322, 2027, 564 and 125 bp) and/or ϕX174 *Hae*III markers (Advanced Biotechnologies) (1353, 1078, 872, 603, 310, 281, 271, 234, 194, 118 and 72 bp), were used to estimate the sizes of DNA fragments. The gel was photographed with a video imaging system. For isolation of small DNA fragments for cloning, 1% (w/v) gels were cast with low-melting agarose in 1x TAE and EtBr (0.5 µg/ml).

22.1.8 Southern analysis

Labelling of the desired probe DNA fragment was done according to Feinberg and Vogelstein (1984). After labelling, the probe was purified by sephadex spin-column chromatography to remove unincorporated nucleotides. The probe was then denatured and pipetted directly into the hybridisation solution.

Approximately 5 µg of genomic DNA were used in each digest and the digests were run out on 0.6% (w/v) 1x TBE agarose gels at 3 V/cm for 6 hours or 1 V/cm overnight. The gel was then capillary blotted onto a Hybond N+ nylon membrane (Amersham) via alkaline transfer in alkaline transfer solution (1.5 M NaCl, 0.25 M NaOH) overnight. The filter was neutralised in a solution of 0.2 M Tris-HCl, (pH 8), 2x SSC and prehybridised in 20 ml of Church-Gilbert buffer (0.5 M sodium phosphate (pH 7.2), 1% (w/v) BSA, 1 mM Na_2EDTA, 7% (w/v) SDS, see Church and Gilbert 1984) at 65°C for a minimum of 4 hours. The prehybridisation buffer was discarded and 15 ml of freshly filtered Church-Gilbert buffer added together with the denatured radioactive probe and hybridised overnight at 65°C. The filter was then washed serially at 65°C in pre-warmed solutions of SSC: 0.1% (w/v) SDS in which the stringency of wash was increased by lowering the concentration of SSC. Typically, washes of 2x, 0.5x and 0.1x SSC were performed. After the final wash, the filter was wrapped in Saran wrap and autoradiographic film was exposed to the filter in an X-ray cassette for 1-7 days at -80°C.

To re-probe filters they were first stripped of radioactive probe by washing at 65°C for a minimum of 2 hours in pre-warmed filter stripping solution (2 mM Tris-HCl (pH 7.5), 0.1% (w/v) SDS, 1 mM Na_2EDTA).

22.1.10 DNA sequencing

DNA was sequenced using the automated services provided by PNACL, University of Leicester and Lark Technologies.

22.2 Extraction of genomic DNA.

Extraction of genomic DNA from a single adult *M. domestica* was carried out according to the protocol for *Drosophila* as described by Hamilton *et al.*, 1991.

Larger scale genomic extractions were carried out as follows: approximately ten adults were frozen in liquid nitrogen, to which 5 ml of homogenisation buffer (160 mM sucrose, 80 mM EDTA and 100 mM Tris pH 8) was added. The flies were then homogenised using a polytron electric homogeniser in 6, 10 second pulses, with 20 second rest intervals on ice. RNaseA was then added to 0.1 mg/ml, with incubation at 37 °C for 30 minutes. SDS to 1% (w/v) and proteinase K to 0.08 mg/ml were then added, with incubation at 50°C for 4 hours. The homogenate was then extracted with equal volumes of phenol-chloroform and then chloroform. The phases were mixed gently and separated using Phase Lock Gel tubes (Flowgen). The DNA was precipitated using an equal volume of ethanol, and sodium acetate to 0.3 M. The DNA was then washed in 70% ethanol and air dried before being resuspended in 0.5 ml of TE.

22.3 DNA amplification by the polymerase chain reaction.

Reactions were carried out in 25µl or 50 µl volumes using 50-100 ng of template DNA. PCR buffer was prepared as described in Jeffreys *et al.*, 1990 (as an 11.1X concentrate). Alternatively, the Expand High Fidelity PCR System (Roche) was used according to the manufacturers instructions. A standard primer concentration of 300 nM was used, which was increased to appropriate levels

when degenerate primers were used. Reaction conditions such as annealing temperature and $MgCl_2$ concentration varied with the primers and template DNA used.

22.4 Construction of suppression-PCR libraries

Suppression-PCR libraries were generated from *M. domestica, L. sericata and C. vicina* genomic DNA and were used to walk both 5′ and 3′ into regions of unknown sequence using PCR with an adaptor primer and gene specific primers (Siebert *et al.*, 1995; Devon *et al.*, 1995; Padegimas and Reichert 1998). Typically 5 µg of genomic DNA was restricted with either blunt cutting enzymes, or sticky ended cutters followed by Klenow mediated end-filling reactions (Sambrook *et al.*, 1989). Agarose gel electrophoresis and Southern transfer of approximately 4 µg of restricted DNA allowed estimates of the size of the fragment of interest and the average fragment size. This allowed calculation of the number of DNA ends to enable efficient adaptor ligation. Adaptor was made by coincidental annealing and phosphorylation of oligonucleotides ol992 and ol993 at 37°C for 1 hour (100 pmol ol992, 100 pmol ol993, 1X PNK forward buffer, 2 mM ATP and 40 units of PNK). The PNK was then denatured at 65°C for 20 minutes, before the adaptor was alcohol precipitated and resuspended in TE to give a concentration of 2 µM. Adaptor was then ligated to each genomic restriction in a ten-fold excess to the approximate concentration of genomic DNA ends, over night at 16°C. The ligations were then diluted 100 fold and 1 µl of these libraries was sufficient template for PCR. sPCR

This method uses PCR primers designed in known sequences to walk into unknown regions of DNA. Degenerate primers can be designed within a conserved domain to first isolate part of the gene and then sPCR can be used to clone and sequence the rest of the gene. The potential for PCR artefacts with the AOL995 primer is high. To eliminate these artifacts a high annealing temperature was used (+60°) and then a second round of PCR was carried out with a nested primer.

2.2.5 Library screening

A *M. domestica* genomic library was screened with a *tll* probe this was carried out according to the protocol in the Promega protocols and applications guide.

2.2.5.1 Estimating the library titre

Seven 10-fold dilutions of the library were made into SM buffer (5.8g NaCl, 2.0g $MgSO_4.7H_2O$, 50ml 1M Tris (pH 7.5) 5ml Gelatin, H_2O to 100ml). 10µl of each dilution was added to 100µl of MRA bacterial cells (O.D. 0.5-0.7) which had been picked from a single colony and grown up in 10ml of LB with 2% maltose and 0.1M $MgSO_4$. This was left at 37° C for 20 minutes and then mixed with 3ml of top agarose (0.7% agar, 0.2% maltose and 10mM $MgSO_4$) at 55° C and poured onto a pre-warmed plate of bottom agar (1.5%). This was left over night at 37° C and the number of plaques was estimated the next day by averaging the results from each dilution.

2.2.5.2 Screening the library

An aliquot of the phage solution (enough for 250,000 plaques per 20x20cm plate) was added to 2ml of MRA bacterial cells and left at 37° C for 20 minutes then mixed with 35ml of top agarose and plated out. This was left at 37° C overnight and then at 16° C for an hour to harden the surface. A Hybond-N nylon membrane (Amersham) was first placed on the surface of the agarose and left for up to 1 min. The filter was then soaked in denaturing solution (1.5 M NaCl, 0.25 M NaOH) followed by netralising solution (0.2 M Tris-HCl, pH 8) and then 2x SSC for 5 mins at a time. A replica filter was left on the surface of the agarose for up to 5 mins before being soaked in the three solutions. The filters were air dried and then exposed to UV radiation to cross-link the phage DNA to the filter.

The filters were probed using the method described for southern blotting (see 2.2.1.8). The resulting x-rays were compared for both filters and if a plaque was visible in the same place on both filters this was deemed to be a positive. Positive plaques were picked from the surface of the plate and resuspended by rotation for 1 hour in SM buffer and 1% Chloroform. These phage suspensions

were rescreened twice to confirm that they contained the probe sequence. In the secondary and tertiary screens a lower concentraton of phage was plated to ensure the plaques were well spaced and this allowed for single colonies to be picked with confidence.

2.2.5.3 Extracting phage λ DNA

Before extracting the DNA the titre of the positive phage had to be increased to more than 10^9 pfu/ml. To do this the phage were plated onto two 90mm plates at a concentration near to confluence. To elute the phage from the plate surface the plate was rotated for 2 hours at room temperature with 1ml of SM buffer. The resulting phage solution was of a high titre and suitable for the DNA extraction procedure.

For the DNA extraction two 140mm plates were prepared with NZY agarose (LB, 1% Casein enzymatic hydrolysate, 0.5% yeast extract, 0.5% Nacl, 0.2% $MgSO_4.7H_20$, 1.5% agarose). The plaques needed to be at near confluence which for 140mm plates was between $5x10^5$ - $1x10^6$ pfu. The appropriate amount of the phage solution was added to 8ml of 0.7% top NZY agarose. The plate was left overnight at 37° C and then rotated for 2 hours at room temperature with 12ml of SM buffer (no gelatin). The eluate was spun at 8,000g for 10 mins at 4° C. The supernatant lysate was treated for 30 mins at 37° C with 1μg/ml *RNase*A and 1μg/ml *DNase*I. An equal volume of 2x NaCl/PEG solution was added to the supernatant and left on ice for 1 hour (50ml of 2x solution: 5.8g NaCl and 9.3g PEG 8000 grade). The solution was spun at 10,000g for 20 mins at 4° C and then the supernatant removed. The pellet was resuspended in 5ml of TE (2.5ml per plate). The solution was spun at 8,000g for 2 mins at 4° C. 20% SDS and 0.5M EDTA pH8 were added to the supernatant at 50μl per 10ml of lysate. This was incubated at 68° C for 15 mins. The solution was then phenol/chloroform extracted. The resulting solution was mixed with 0.2M NaCl and 2 volumes of isopropanol. This was left on ice for 1 hour and then spun at 12,000g for 30 mins at 4° C. The pellet was washed with 5ml of 70% ethanol and spun at 12,000g for 10 mins at 4° C. The pellet was air dried

and than resuspended in 200µl TE. The concentration of λ DNA was quantified by standard methods.

The *M. domestica* genomic DNA was excised from the λ arms using *Sau*3AI. The resulting 15kb genomic fragment was digested with *Eco*RI and subcloned into pbluescript. The 15kb fragment was mapped using the restriction enzymes *Bgl*II, *Eco*RI, *Eco*RV, *Hind*III, *Pst*I, *Xba*I.

22.6 mRNA extraction

mRNA was extracted from *M. domestica* early embryos using a Stratagene mRNA isolation kit according to the protocols supplied therein. The mRNA concentration of extracts was estimated by comparing the fluorescence of a serial dilution in EtBr, to that of known concentrations of yeast tRNA.

22.7 5′ and 3′ Rapid Amplification of cDNA Ends (RACE) - PCR

5 RACE-PCR was performed using the Gibco BRL 5′ RACE System Version 2 and the protocols supplied therein. This method allows the cloning of the 5′ end of a specific transcript from a short stretch of known downstream sequence. Basically, a cDNA is synthesised from an mRNA template using a gene specific primer and Reverse Transcriptase (RT). The cDNA then has a cytosine rich tag added using TdT (Terminal deoxynucleotidyl Transferase) and this allows the use of PCR to amplify a specific product using a primer based on the tag sequence (AAP) and a nested gene specific primer.

3′ RACE PCR was performed using the Gibco BRL 3′ RACE System according to the manufacturers instructions. Using this method the 3′ end of a transcript can be cloned based on primers designed in known upstream sequence. Adaptor primer (AP) is annealed to the poly A tails of an mRNA population and cDNAs are then generated using RT. A gene specific primer is then used in combination with another primer based on the AP sequence (AUAP) to amplify a specific product using PCR.

22.8 *DNase*I footprinting

*DNase*I footprinting was carried out according to a standard protocol (Galas and Schmitz, 1978; Lin and Shiuan, 1995).

22.8.1 Primer end-labelling

Primers were end-labelled with [^{33}P] for 30 minutes at 37°C in the following reaction (10 µl): 10 pmol primer, 0.5 µl T4 PNK, (10 units/ul), 5 µl [^{33}P] y-ATP, (111 TBq/mmol), 1x PNK forward reaction buffer and water. The reaction was stopped by heating to 65°C for 15 minutes.

22.8.2 PCR

Labelled primers were used to generate labelled PCR probes in 50 µl reactions of the following composition: 2 – 5 ng of plasmid template, end labelled primer at 0.1 µM, opposing primer at 0.1 µM, 1.5 mM $MgCl_2$, 1X React IV PCR buffer, 0.2 mM dNTPs and 0.5 µl *Taq* polymerase (Advanced Biotechnologies). PCR was carried out for 22-25 cycles under appropriate conditions. The product was purified using a Qiagen PCR purification kit and then quantified on a minigel.

22.8.3 Protein synthesis

M domestica and *D. melanogaster* Bcd homeodomain-GST fusion proteins were synthesised from pBCDR1 and PBcdTN3 (table 2.1) respectively by P. Shaw, using the method described in McGregor *et al.*, 2001. Concentrations of active protein were estimated by gel-shift assays using the method described in Zhao et al. 2000.

22.8.4 Binding reaction and *DNase*I digestion

Labelled PCR probe was incubated with protein for 30 minutes at room temperature in the following binding reactions (50 µl): 10 ng DNA, protein (at 100 nM, 10 nM or 1 nM), 100 ng dI:dC and 25 µl of 2X binding buffer (80 mM Tris pH 7.5, 0.2 M NaCl, 40% glycerol, 0.2% Triton X-100, 2mM DTT). In the control reactions either no protein was added or GST tag was added to 0.2 µg/ml. 50 µl

of 50 mM $MgCl_2$/10 mM $CaCl_2$ was then added and the reactions placed on ice. *DNaseI* was then added to a final concentration of 0.75 µg/ml and the reactions incubated on ice for 5 minutes. Digestion was stopped by the addition of 90 µl of stop mix (0.1 M EDTA, 1% SDS, 0.2 M NaCl and 0.1 mg/ml yeast tRNA). The reactions were then extracted with an equal volume of phenol/chloroform and the DNA precipitated with two volumes of 100% ethanol. The pellet was washed in 70% ethanol, air dried and resuspended in 3 µl of sequencing gel loading buffer.

22.8.5 DNA sequencing

Plasmid DNA obtained from Qiagen minipreps was denatured by incubation in 0.2 M NaOH, 0.1 mM Na_2EDTA for 20 minutes at 37°C. Denatured DNA was precipitated with ethanol and resuspended in 10 µl of TE. 1-3 µg of denatured plasmid were used per sequencing reaction. 1-3 pmol of sequencing oligonucleotide were annealed to 1-3 µg of denatured double-stranded DNA by heating to 70°C for 3 minutes and cooling slowly to 45°C (1°C/min) in sequencing buffer (40 mM Tris-HCl (pH 7.5), 20 mM $MgCl_2$, 50 mM NaCl). Labelling and termination reactions were done as described in the Sequenase v2.0 protocol (Amersham). Termination mixes were made according to the T7 sequencing kit (Pharmacia Biotech.). Termination reactions were done in microtitre plates for 4 minutes at 37°C. Samples were denatured by heating for 2 minutes at 80°C just before loading on to gels.

22.1.8 Denaturing polyacrylamide (sequencing) electrophoresis

Glass plates 21 x 50 cm from a Sequi-Gen sequencing gel apparatus set (Biorad) were used. 5% polyacrylamide gels were cast using 'Sequagel' (gas-stabilised 19:1 acrylamide:bisacrylamide acrylamide solution in 8.3 M urea, National Diagnostics, Flowgen), in 1x TBE, according to the manufacturer's recommendations. 24, 0.4 mm thick teflon sharkstooth combs (Biorad) were used to make the wells for sample loading.

Gels were run as 'gradient' gels, the top buffer was 0.5x TBE and the bottom buffer 1x TBE. After samples had entered the gel (10-15 minutes after

loading), 1/2 the volume of the bottom buffer of 3 M sodium acetate was added to the bottom buffer, which lowers the conductivity of the lower buffer establishing an ionic gradient which creates a more linear rate of migration for the smaller fragments. After the gel run was complete, gels were fixed in a solution of 10% (v/v) acetic acid, 15% (v/v) methanol for 10 minutes. Gels were dried onto Whatman 3MM paper in a vacuum drier (Biorad model 583) at 80°C for 60-90 minutes. Gels were exposed to X-ray film (Fuji RX100) for 1-7 days at room temperature.

Regions of protected sequence were distinguished by the comparison of digestion patterns between samples and controls. When a protein binds to DNA this can cause the DNA to bend and increase the exposure of nearby sequences to DNaseI. This effect is seen as hypersensitive bands on gels.

22.9 *In situ* hybridisation of whole-mount embryos

Whole mount in situ hybridisations were carried out on *M. domestica* and *D. melanogaster* embryos essentially as described by Tautz and Pfeifle (1989), with modifications (Bonneton *et al.*, 1996).

22.9.1 *In vitro* transcription for synthesis of riboprobes

Approximately 2 µg of Qiagen-purified plasmid DNA containing a cloned insert of DNA was linearised by restriction digestion. After completion of the digestion the linearised plasmid was purified using a column from a PCR purification kit (Qiagen) and eluted in 30 µl of DEPC-treated (RNase-free) water. Linearised plasmid was then used as template for *in vitro* transcription using DIG DNA-labelling kit components (Roche) in the following reaction: linearised template, 10 µl; 1x component buffer (40 mM Tris-HCl pH 8.0, 6 mM $MgCl_2$, 10 mM DTT, 2 mM Spermidine, 10 mM NaCl, 0.1 units RNase inhibitor); 1x rNTPs (1 mM rATP, rGTP, rCTP, 0.65 mM DIG-11-UTP); RNasin (Promega), 20 units and T7 or T3 RNA polymerase, 40 units. Transcription was performed at 37°C for 2 hours and was stopped by heating to 65°C for 10 minutes to inactivate the enzyme. RNA was precipitated with LiCl and ethanol to remove unused rNTPs and

resuspended in 50 µl DEPC-treated water with 20 units of RNasin (Promega). The yield and integrity of product was tested by agarose gel electrophoresis and was typically about 8 µg of riboprobe per reaction.

22.9.2 Dechorionation

Cat meat was placed on petri dishes to collect embryos from *M. domestica* and apple juice plates were used for *D. melanogaster* (10.75 g of agar dissolved in 237 ml of distilled water, 5 ml acid mix [per litre: 41.5 ml phosphoric acid, 418 ml propionic acid, 30 ml food colouring] and 245 ml of apple juice). The embryos were removed with a brush and transferred to a wire basket, where they were rinsed with distilled water and dechorionated with household bleach (about 5% (w/v) $Na(HClO)_3$) in a watch-glass for two minutes. The embryos were then rinsed thoroughly with water to remove the bleach.

22.9.3 Fixation

Dechorionated embryos were fixed in screw-capped glass vials containing 1.82 ml of DIG-FIX solution, 2 ml heptane, 0.68 ml formaldehyde (~37% solution, stabilised with 10-15% (v/v) methanol, Sigma). Vials were placed on a rotating wheel for 30 minutes at room temperature. Fixed embryos were aspirated from the organic-aqueous interface with pasteur pipette and transferred to a fresh vial containing 2 ml methanol and 1 ml heptane. The vitelline membrane was removed by vortexing on the lowest setting on an electric vortex for 30-60 seconds. De-vitellinised embryos sink into the methanol layer and were collected by aspiration with a pasteur pipette. The embryos were washed once with 1 ml of methanol to ensure dehydration and stored in methanol at -20°C.

22.9.4 Pre-treatment of embryos for *in situ* hybridisation

All washes and incubations described in this and subsequent steps carried out in 1 ml volumes of liquid on a rotating wheel at room temperature unless otherwise stated. Fixed embryos were re-hydrated by washing for 3 minutes in

methanol:PBT (1:1) then twice in PBT (phosphate buffered saline with tween; 130 mM NaCl, 70 mM Na_2HPO_4, 30 mM NaH_2PO_4 and 0.1% (w/v) Tween). Rehydrated embryos were post-fixed for 20 minutes in PBT: 5% formaldehyde. Post-fixation was stopped by rinsing embryos briefly in PBT, then embryos were washed twice for 5 minutes each in PBT. The *M. domestica* embryos were then washed for 5 minutes in PBT plus 5 µg of proteinase K. Proteinase K digestion was stopped by rinsing briefly in PBT and then washing twice for 5 minutes in PBT plus 2 mg of glycine. The embryos were then post-fixed a second time for 20 minutes in PBT; 5% (v/v) formaldehyde which was stopped by rinsing briefly in PBT and then washing twice for 5 minutes in PBT.

22.9.5 Pre-hybridisation and hybridisation

Pre-treated embryos were washed in 0.5 ml of a 1:1 solution of PBT: Hyb-D (50% (v/v) deionised formamide, 5x SSC, 0.1% (w/v) Tween-20, 1 mg/ml yeast tRNA, 2% (w/v) DIG blocking reagent) for 15 minutes. The embryos were then transferred to 0.5 ml of Hyb-D and incubated for 30 minutes at 55°C, then 65°C for 1 hour to denature endogenous enzymes. The embryos were then returned to 55°C and incubated for 30 minutes. Embryos were hybridised in 0.1 ml of fresh Hyb-D together with 10-50 ng of riboprobe overnight at 55°C.

22.9.6 Pre-immunoreaction and immunoreaction

Embryos were washed to remove unbound riboprobe. Washes in 0.5 ml solution of 4:1, 3:2, 2:4. 1:4 Hyb-D:NTB (150 mM NaCl, 100 mM Tris-HCl (pH 7.5), 0.1% (w/v) Tween-20, 0.2% (w/v) DIG blocking reagent) were carried out for 10 minutes each at 60°C. After a final wash for 10 minutes in 0.5 ml NTB at 60°C, embryos were pre-incubated for 4 hours in 1 ml NTB plus 2% (w/v) goat serum (Boehringer Mannheim) at 4°C.

Embryos were incubated overnight at 4°C in 1 ml NTB, 2% (v/v) goat serum, Anti-DIG-AP antibody (polyclonal Fab fragments conjugated to alkaline phosphatase, Boehringer Mannheim) diluted 1/2000.

22.9.7 Colour staining

Embryos were washed three times in PBT for 30 minutes each at 4°C, then washed twice for 10 minutes each at 4°C and finally once at room temperature in 1 ml of colouration solution (CS; 0.1 M NaCl, 0.1 M Tris-HCl (pH 9.5), 50 mM $MgCl_2$, 0.1% (w/v) Tween-20). Colouration reaction was initiated by incubating the embryos in 1 ml CS + 0.45 µl NBT + 3.5 µl X-phosphate (NBT from the DIG kit: nitroblue tetrazolium salt dissolved in 70% (v/v) dimethylformamide). Staining was checked after 2 hours and stopped after 3-4 hours by washing embryos twice in 1 ml PBT. Embryos were dehydrated by rinsing in 1 ml ethanol:PBT (1:1) then twice in 1 ml absolute ethanol.

22.9.8 Permanent mounting, microscopy and photography

Dehydrated embryos were washed in 0.5 ml 1:1 ethanol:Spurr (Spurr: low viscosity embedding medium (hard composition), Sigma) and then 0.3 ml of Spurr. The embryos were left to settle to the bottom of the tube and were taken up in a small volume of Spurr (about 60 µl) and mounted on a glass slide. Slides were incubated overnight at 65°C and analysed using a Nikon Optiphot-2 microscope. Photographs were taken at 200x magnification with a Nikon exposure unit with automatic exposure times, typically 60-120 ms.

22.10 *In vitro* DNA binding assays

Bcd-GST fusion proteins for in vitro DNA binding experiments were expressed and purified as described previously (see 2.2.8.3; McGregor et al. 2001).

2.2.10.1 Binding reaction

33P end-labelled primers were used in PCRs to generate labelled promoter fragments containing Bcd binding sites. Approximately 0.5 ng of labelled DNA were titrated with increasing amounts of Bcd protein; the reactions were bound at room temperature for 30 mins in binding buffer (20 mM Tris pH 7.5, 40 mM NaCl, 0.5 mM EDTA, 20% glycerol, 1 mM DTT, 0.1% Triton X-100) and run on 4% polyacrylamide gels at 5V / cm. The running buffer used was 285 mM glycine,

37.5 mM Tris, 0.15 mM EDTA. The gels were dried and exposed to X-ray film. The following promoter regions were analysed: *Drosophila tll* 4715-4883 of EMBL database #AF019362 (Bcd binding regions 4-8; Liaw and Lengyel 1992;), *Musca tll* 528-717 of EMBL database #AJ421995 (this work, Bcd binding regions 10 - 13, figure 5).

2.2.10.2 Quantitative analysis of DNA binding data

The fraction of DNA bound (in all complexes) in gel-shift experiments was determined by comparing band intensities of bound and free DNA complexes. The mean data from the gel-shift experiments (four replicates for each interaction, fitting done with weighting against standard deviations for each datapoint) were fitted to the equation using the program DATAFIT (Oakdale Engineering, Portland, USA): Ybar = Kn*Xn / (1 + Kn*Xn). Where Ybar is the saturation (fractional occupancy in footprinting, fraction DNA bound in gel-shifts). K is the apparent equilibrium constant, X is the concentration of Bcd protein and n is the Hill coefficient. The affinity constant (Km) is the reciprocal of the equilibrium constant. The Hill coefficient n describes the cooperativity of the system, where values greater than 1 indicate positive cooperativity.

22.11 Creating a *D. melanogaster* transformant fly

22.11.1 Consructing the injection plasmid

The plasmid was based on the PCaSpeR AUG βgal vector of Thummel *et al.*, 1988. The 2.2kb *tll* promoter fragment was amplified with primers containing *Kpn*I and *Bam*HI sites in a PCR reaction with the Expand High Fidelity PCR System (Roche) used according to the manufacturers instructions. This fragment was cloned into TOPO vector and sequenced. The plasmid was digested with *Kpn*I/*Bam*HI and cloned into the PCaSpeR AUG βgal vector using the *Kpn*I and *Bam*HI sequences in the multiple cloning site. The resulting plasmid was Maxiprepped (QIAGEN) and resuspended in injecting buffer (0.5 KCl, 10mM Na_2HPO_4 (pH 6.8) for 100x solution).

22.11.2 Injecting the plasmid

Germline transformation was performed using the standard protocol (Rubin and Spradling, 1982). Stocks of SbeΔ2-3 *D. melanogaster* were put onto apple juice egg laying plates and left for 30 mins (Robertson *et al.*, 1988). These eggs were discarded and a fresh collection was made after another 30 mins. The eggs were dechorionated by hand and aligned on a microscope slide using double sided tape. The embryos were dessicated for approx 5-8 mins and then covered with a small drop of Voltalef oil (grade 10S). Glass microcapillary tubes were pulled into fine pointed needles using a micropipette puller by H Rowe. The needle was mounted in a Narishige micromanipulator attached to a nitrogen pneumatic pico pump and an open bevelled edge made by breaking the end with a scalpel blade. The plasmid DNA (0.5 -0.75μg/ml) was injected into the posterior of each embryo. Any old or leaking embryos were destroyed before the tape was placed onto *D. melanogaster* grape food plates. Larvae were collected from the plates and transferred to vials containing sugar food.

22.11.3 Identifying transgenic flies

The injected adult flies were crossed with w118 *D. melanogaster* stocks and the F1 generation checked for red eyes. Red eyed flies were back-crossed to w118 stocks and in the F2 generation Sb flies removed. This removed the P-element from the line and stabilised the insertion. To map the insertion to a chromosome males of each F1 red eyed line w also crossed to w;ScO/Cyo;MKRS/TM6 females and individual red eyed, CyO, F2 flies crossed to w118. If there were no red eyed F2 males then the insert mapped to the X chromosome. If individual flies of the F3 generation were both red eyed and CyO then the insert mapped to the third chromosome, if such individuals were absent then the insert mapped to the second chromosome.

22.11.4 Inverse PCR to map insertions

To confirm the insertions in each transgenic line were independent of each other inverse PCR was used. The technique followed was that published on the BDGP website under resources/methods (http://www.fruitfly.org).

22.12 Computer analysis

Sequence alignments were made using the Clustal W program (Thompson *et al.*, 1994), the GCG algorithm PILEUP or DiAlign for the less conserved sequences (http://bibiserv.techfak.uni-bielefeld.de/dialign/). Dotplots were generated using COMPARE and DOTPLOT; binding sites were predicted using FINDPATTERNS; all of these programs are available on the GCG (1994) package, version 8.1.

Consensus sequences for the Bcd-binding sites in the *tll* promoters were calculated from these alignments by the frequency of bases at each position. If a position did not have one particular base present in 50% or more of the aligned sequences then that position was left ambiguous in the consensus sequence.

PEPTIDESORT (also in GCG) was used to predict the molecular weights of the Bcd proteins from each species. NIH Image 1.61 was used to compare the density of hybridising bands in the band shift assays. The SIMPLE 34 program (Hancock and Armstrong 1994) was used in the analysis of sequence simplicity and this is explained in full in 5.2.1. The *cad* mRNA secondary structures were predicted using the Mfold programme available on the Mfold website (http://www.bioinfo.rpi.edu/applications/mfold/).

Chapter 3 Characterisation of the *tll* gene in *M. domestica*

3.1 Introduction

3.1.1 Expanding the evolutionary study of Bcd regulation

The interaction between Bcd and the *hb* promoter has been compared between *D. melanogaster* and *M. domestica*. To expand this comparison to other Bcd regulated promoters, additional *M. domestica* genes were cloned and sequenced (this work; McGregor, 2002). This chapter describes the isolation of *M. domestica tll* and evolutionary analysis of both the coding sequences and mRNA expression patterns.

3.1.2 Tll protein structure

First identified in *D. melanogaster, tll* encodes a transcription factor containing a zinc finger DNA binding domain and is a member of the steroid receptor superfamily (Strecker *et. al.*, 1986; Pignoni *et. al.*, 1990). The zinc finger is highly conserved between all species so far examined even over large phylogenetic distances; for example, the zinc finger of *D. melanogaster* Tll is 81% conserved with that of chick, Tlx (Yu et al, 1994). The Tll protein also contains a motif immediately C-terminal to the DNA-binding domain, called the T/A box, which is thought to function in DNA sequence recognition and dimerisation. This T/A box is also highly conserved between *tll* orthologues (Yu *et al.*, 1994; see fig. 3.1a).

In addition, the Tll protein contains a C-terminal ligand-binding domain that is much less conserved than the DNA binding domain and an associated ligand has yet to be identified (Pignoni *et al.*, 1990). Within, this domain is a PEST motif that marks the protein for degradation (Rogers *et al.*, 1986). The presence of a PEST motif is typical of early developmental transcription factors present at the syncytial blastoderm stage, when RNA and proteins are free to diffuse. Persistance of a transcription factor into the cellular stages of development may cause problems with the ectopic activation or repression of target genes (Irish *et al.*, 1989; Lall *et al.*, 2003).

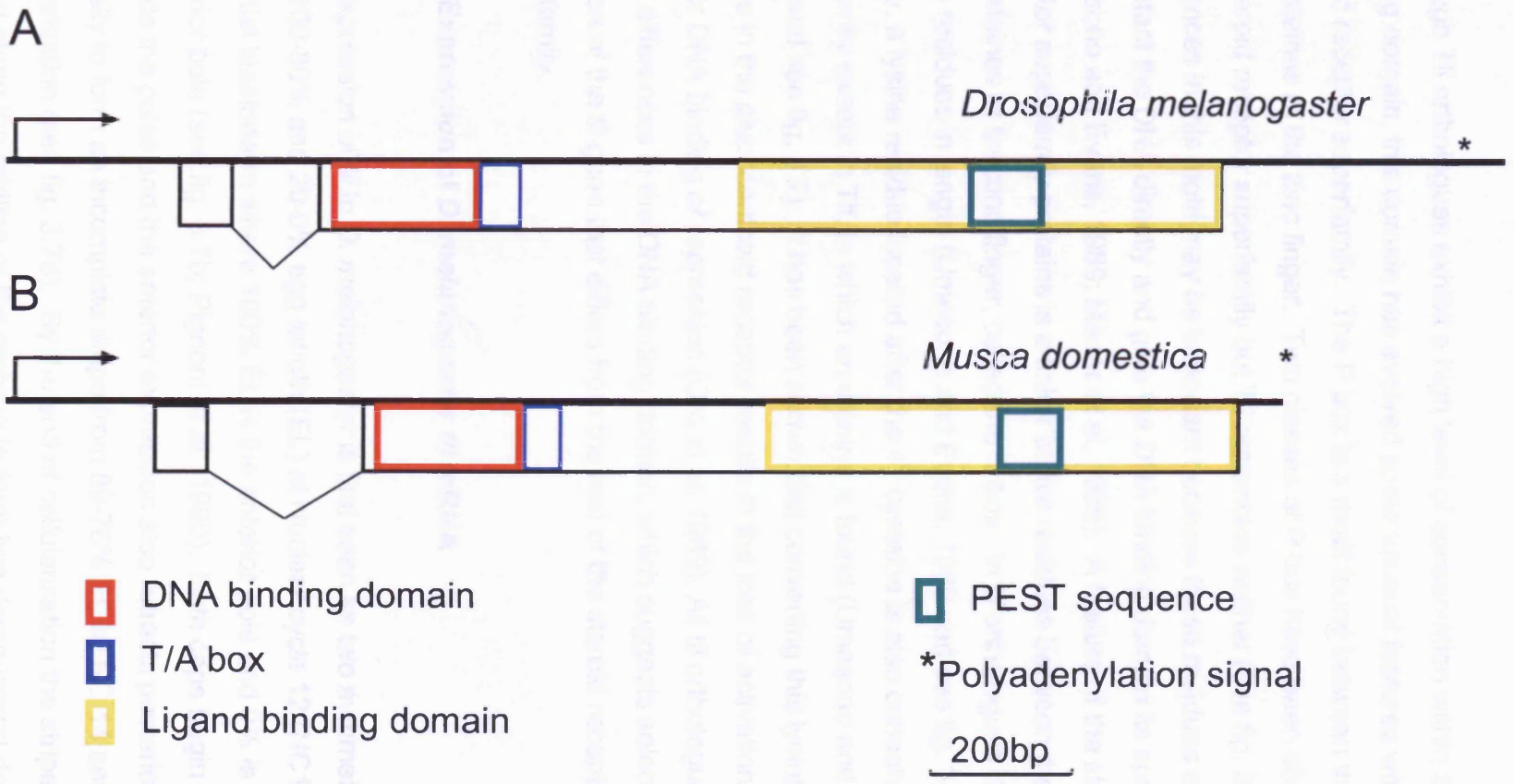


Figure 3.1 Comparison of the *M. domestica tll* transcript structure with that of *D. melanogaster*.
A. *D. melanogaster tll* transcript structure. B. *M. domestica tll* transcript structure. The arrow indicates the transcription start site, the open black boxes are the coding regions, coloured boxes indicate the functional domains (see key).

3.1.3 Tll evolution within the steroid receptor superfamily

Although Tll orthologues exhibit a high level of conservation within the DNA binding domain, this domain has evolved some unusual features within the steroid receptor superfamily. The P box is a motif found between the 3rd and 4th cysteines of the zinc finger. Two classes of P box have been observed in the steroid receptor superfamily but Tll resembles neither (see fig. 3.2). The differences in this motif may be important because these residues are thought to contact the DNA directly and give the DNA binding domain its specificity (Umesono and Evans, 1989; Mader et al, 1989). A feature of the steroid receptor superfamily proteins is a linker of five residues between the 5th and 6th cysteines of the zinc finger, called the D box. In *tll* orthologues this linker is seven residues in length (Umesono and Evans, 1989 and see fig. 3.2). Finally, a lysine residue located after the 4th cysteine is also conserved within this family except in Tll, in which an alanine is found (Umesono and Evans, 1989 and see fig. 3.2). It has been shown that converting this lysine to a glycine in the glucocorticoid receptor results in the loss of activation potential but not DNA binding or repression (Oro et al, 1989). All *tll* orthologues exhibit these differences in the DNA binding domain, which suggests selection for a function of the *tll* gene that differs from the rest of the steroid receptor superfamily.

3.1.4 Expression of *D. melanogaster tll* mRNA

The expression of *tll* in *D. melanogaster* is first seen as two symmetrical caps from 100-80% and 20-0% egg length (EL) at Nuclear cycle 12 (NC12) in the syncytial blastoderm where 100% EL is the anterior pole and 0% is the posterior pole (see fig. 3.7b; Pignoni *et al.*, 1990). Both caps begin to retract towards the poles and the anterior expression also retracts posteriorly and ventrally to form an incomplete stripe from 89-76% EL by NC 14 and cellularisation (see fig. 3.7d). By the end of cellularisation the stripe has divided along the midline of the embryo to form two dorso-lateral domains. The posterior expression domain recedes to 15-0% EL. As gastrulation begins the posterior expression rapidly disappears and the anterior stripes

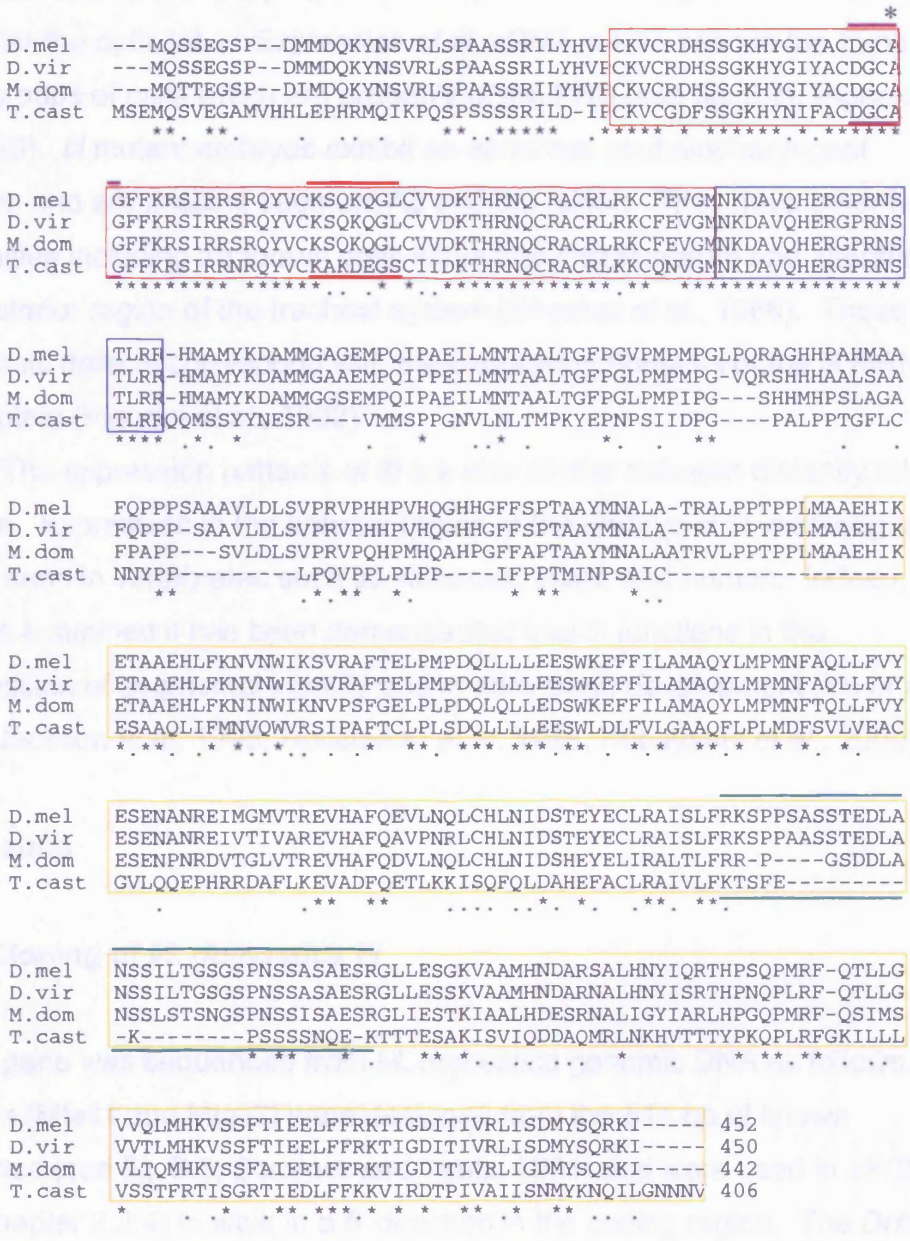


Figure 3.2 TII protein comparison between insect species: D. mel - *Drosophila melanogaster* (Pignoni *et al*, 1990), D. vir - *Drosophila virilis* (Liaw *et al*, 1993), M. dom - *Musca domestica* (this work), T. cast - *Tribolium castaneum* (Schroder *et al*, 1997).
The DNA binding domain is boxed in red, the T/A box in blue and the Ligand binding domain in yellow (see text). The DNA recognition motif within the DNA binding domain is underlined in purple, with the unique alanine marked with an asterisk. The 7 residues between cysteines 5 and 6 are underlined in red and the PEST domain is underlined in green. Numbering indicates the length of the proteins.

become obliquely inclined. During germ band elongation *tll* expression overlaps with the developing brain but later in development the expression becomes restricted to the peripheral and post-cortical regions of the brain, in particular the optic lobe. Expression of *tll* mRNA is also seen in the trunk in small groups of cells which are probably of the PNS (see fig. 3.7f; Pignoni *et al.*, 1990). *tll* mutant embryos exhibit an abnormal cephalopharyngeal skeleton and are missing segment A8 and the telson. They have other head deformities including abnormal optic lobes and clypeolabrum and alteration of the posterior region of the tracheal system (Strecker *et al.*, 1986). These embryonic defects correspond with the expression domains of the *tll* mRNA and protein (Pignoni *et al.*, 1992).

The expression patterns of *tll* are also similar between distantly related species. Expression in the anterior region of the embryo in *D. melanogaster* is also seen in vertebrates such as Xenopus, chick, and human. Indeed, in all species examined it has been demonstrated that *tll* functions in the suppression of segmental identity and in early brain development (Yu *et al*, 1994; Jackson *et al*, 1998; Hollemann *et al*, 1998; Kobayashi *et al.,* 2000).

3.2 Results

3.2.1 Cloning of *M. domestica tll*

The *tll* gene was sequenced from *M. domestica* genomic DNA as follows. Primers (Mtail1 and Mtail2) were designed from the 511 bp of known sequence (see fig. 3.3; Sommer and Tautz, 1991) and were used in sPCR (see chapter 2.2.4) to walk in a 5' direction in the coding region. The *Dra*I and *Ssp*I sPCR libraries gave products of approximately 800 bp and 900 bp respectively. Further sPCR, using primer pairs Mtail3/4 and Mtail5/6 yielded approximately 2.5 kb of sequence 5' to the start of the coding region. Additional primers were designed (Mtail7 to 11) but were unsuccessful under a variety of PCR conditions (see 2.2.3). This may have been due to the presence of long runs of A and T in the target sequence. For a complete list of primers see 2.1.4.

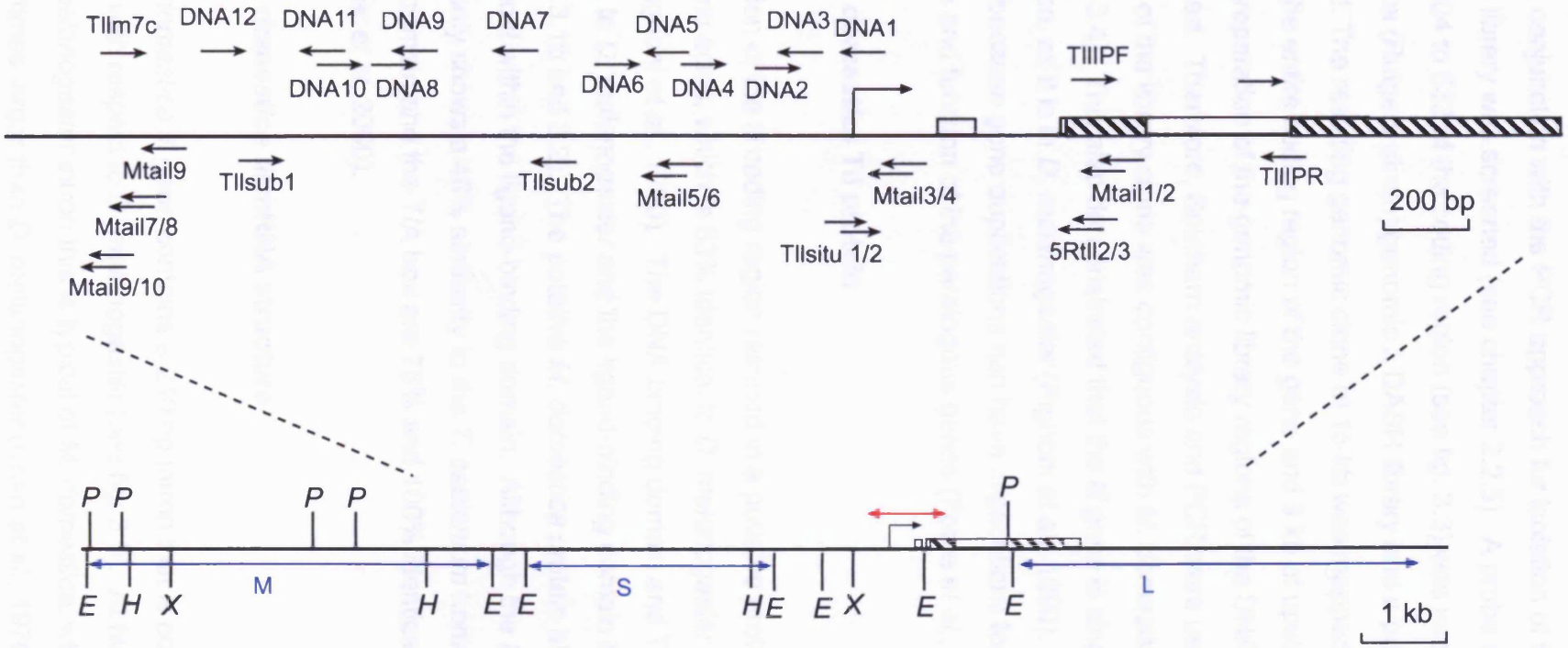


Figure 3.3
Sequencing of *M. domestica tll*.
Primer position and orientation are shown in relation to the *tll* transcript (see text for use and 2.1.4). The arrow indicates the transcription start site and the boxes the coding region (striped boxes represent functional domains). The expanded region is shown within the clone isolated from the *M. domestica* genomic DNA library. Restriction sites are shown in relation to the coding region and probe shown in red; *E* - *Eco*RI; *H* - *Hind*III; *X* - *Xba*I. The *Eco*RI sites represent the boundaries of the fragments that were subcloned into pBluescript for sequencing, the larger fragments were called S, M and L. The primers used to confirm the restriction map are given in table 2.2.

In conjunction with the PCR approach for isolation of the *tll* sequence a genomic library was screened (see chapter 2.2.5). A probe corresponding to bases -304 to 532 of the coding region (see fig. 3.3) was used to screen a *M. domestica* (Rutgers strain) genomic λ DASH library and a positive clone identified. The resulting genomic clone of 15 kb was mapped and found to contain the entire coding region of the gene and 9 kb of upstream sequence. During preparation of the genomic library regions of the DNA can become rearranged. Therefore, Southern analysis and PCR were used to confirm that the map of the library clone was contiguous with *M. domestica* genomic DNA (see fig. 3.4). This also demonstrated that the *tll* gene is single copy in *M. domestica,* as it is in *D. melanogaster* (Pignoni *et al.*, 1990). This is important to know because gene duplications can have implications for the evolution of structure and function of the paralogous genes (Force *et al.*, 1999).

3.2.2 *M. domestica* Tll protein

Translation of the *tll* coding region resulted in a putative protein sequence of 442 amino acids, which is 83% identical to *D. melanogaster* Tll (452 amino acids; Pignoni *et al.,* 1990). The DNA binding domain and T/A box are identical to *D. melanogaster* and the ligand-binding domain is 62% identical (see fig. 3.1b and 3.2). The putative *M. domestica* protein also contains a PEST motif within the ligand-binding domain. Although the *M. domestica* Tll protein only shows a 46% similarity to the *T. castanium* homologue, the DNA binding domain and the T/A box are 78% and 100% identical respectively (Schroder *et al*, 2000).

3.2.3 *M. domestica tll* mRNA structure

The *M. domestica tll* gene contains a 210 bp intron that is conserved in position with respect to *D. melanogaster* (see fig. 3.5). At twice the length of the *D. melanogaster* intron this is typical of *M. domestica*, whose genome is 3.5 to 5 times larger than *D. melanogaster* (Crain *et al.*, 1976). The splice sites are conserved, as is a putative branch site, which is identical to the *Drosophila* consensus sequence (Cavener, 1987 and see fig. 3.5).

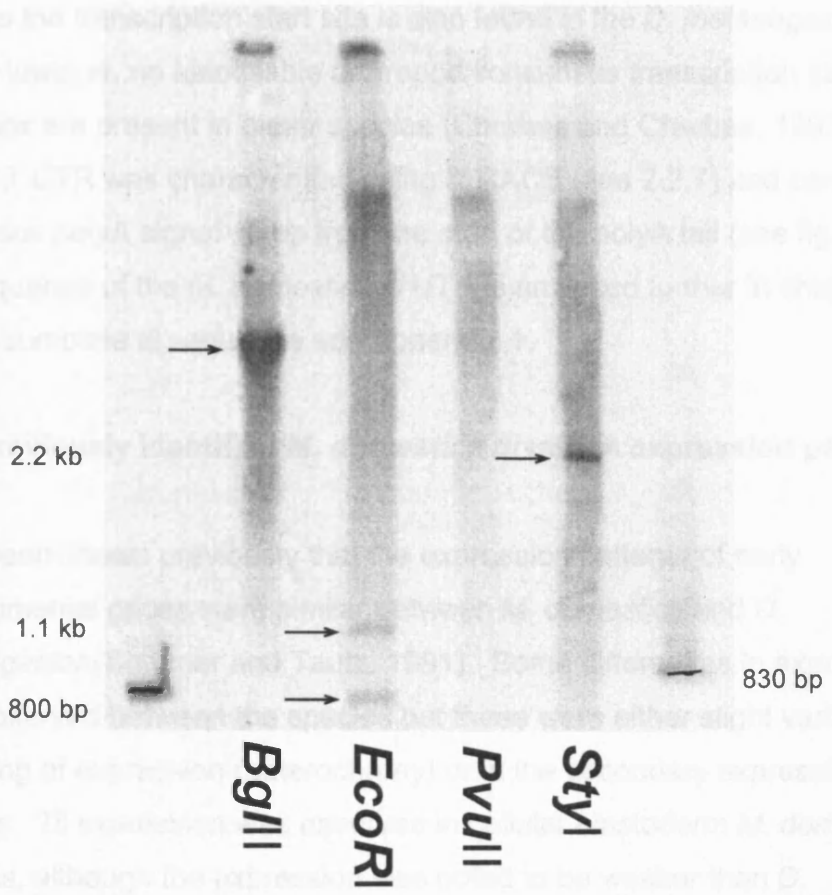


Figure 3.4
Southern blot of *M. domestica* genomic DNA using the probe from the library screen. Enzymes used were *Bgl*II; *Eco*RI; *Pvu*II; *Sty*I. The outside lanes contain free probe at approximately 830 bp in length. An *Eco*RI site is present in the probe and thus two bands are seen in the Southern blot.

The 5' UTR was identified by 5' RACE (see 2.2.7) and is 199 bp in length, in comparison to 232 bp in *D. melanogaster*. A 13 bp sequence that overlaps the transcription start site is also found in the *D. melanogaster tll* gene. However, no identifiable arthropod consensus transcription start site or TATA box are present in either species (Cherbas and Cherbas, 1993). The 365 bp 3' UTR was characterised using 3' RACE (see 2.2.7) and contains a consensus polyA signal 45 bp from the start of the polyA tail (see fig. 3.5). The sequence of the *M. domestica tll* UTR is analysed further in chapter 5. For the complete *tll* sequence see Appendix 1.

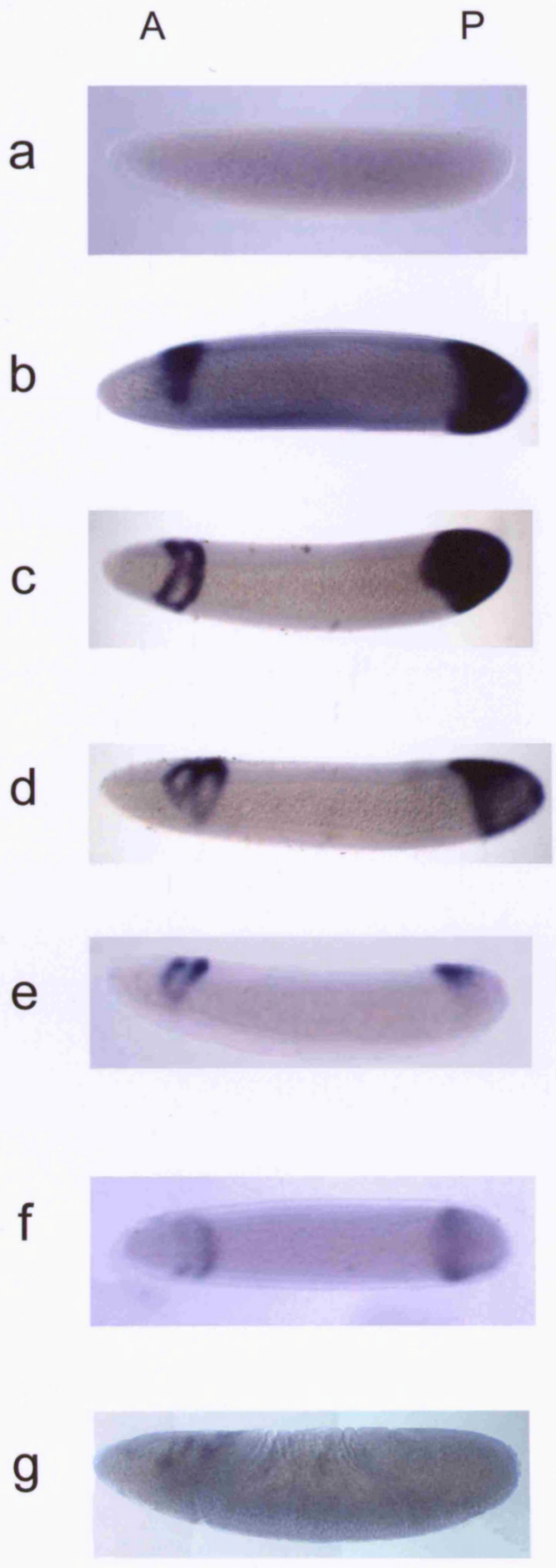
3.2.4 Previously identified *M. domestica tll* mRNA expression patterns

It had been shown previously that the expression patterns of early developmental genes were similar between *M. domestica* and *D. melanogaster* (Sommer and Tautz, 1991). Some differences in expression were observed between the species but these were either slight variations in the timing of expression (heterochrony) or in the secondary expression patterns. *Tll* expression was observed in cellular blastoderm *M. domestica* embryos, although the expression was noted to be weaker than *D. melanogaster* at the same stage (Sommer and Tautz, 1991). *Tll* mRNA was shown to be present in two stripes at the anterior of the embryo and in a single stripe in the posterior with dorsoventral asymmetry (Sommer and Tautz, 1991 and see fig. 3.6c). My study used a different *M. domestica tll* probe and was able to identify further expression patterns at both earlier and later stages.

3.2.5 The complete expression of *tll* mRNA in *M. domestica*

An RNA probe corresponding to -304 to 532 of the *tll* coding region was used for the *in situ* experiments. A sense probe was used in a control experiment but did not result in any staining (data not shown). The first expression of *M. domestica tll* mRNA was seen in the cellular blastoderm, stage 5 (see fig. 3.6b). Staining appears in a posterior cap from 0-19% and an anterior dorso-lateral stripe from 84-75% EL (n=32) and looks very similar to the stripe seen

Figure 3.5
Comparison of the *tll* transcript sequence between *M. domestica* and *D. melanogaster*.
A. Alignment of 5'UTR sequences and the start of the coding region. The black arrow denotes the transcription start site the red thymine is position one of the cDNA. The red asterisks highlight the region of conservation between the two species. The blue arrow denotes the translation start site, with the first ATG highlighted. The *Drosophila* translation consensus sequence present in the *M. domestica* sequence is also highlighted in blue (the *D. melanogaster* sequence does not contain a translation consensus sequence). The other ATG sequences present near the N-terminus are shown in pale blue. The protein sequence for both species is shown below in grey, the methionine residues highlighted in blue.
B. Conservation of the intron position between *M. domestica* and *D. melanogaster*. The intron is in italics and separated from the coding region by black bars. The pink residues denote invariant bases at the start of the intron and within the putative branch point, which is underlined. The diagonal lines mark gaps in the sequence and the grey residues denote the Tll protein sequence. The *Drosophila* intron consensus sequence is shown in blue.


A
P
a
b
c
d
e
f
g

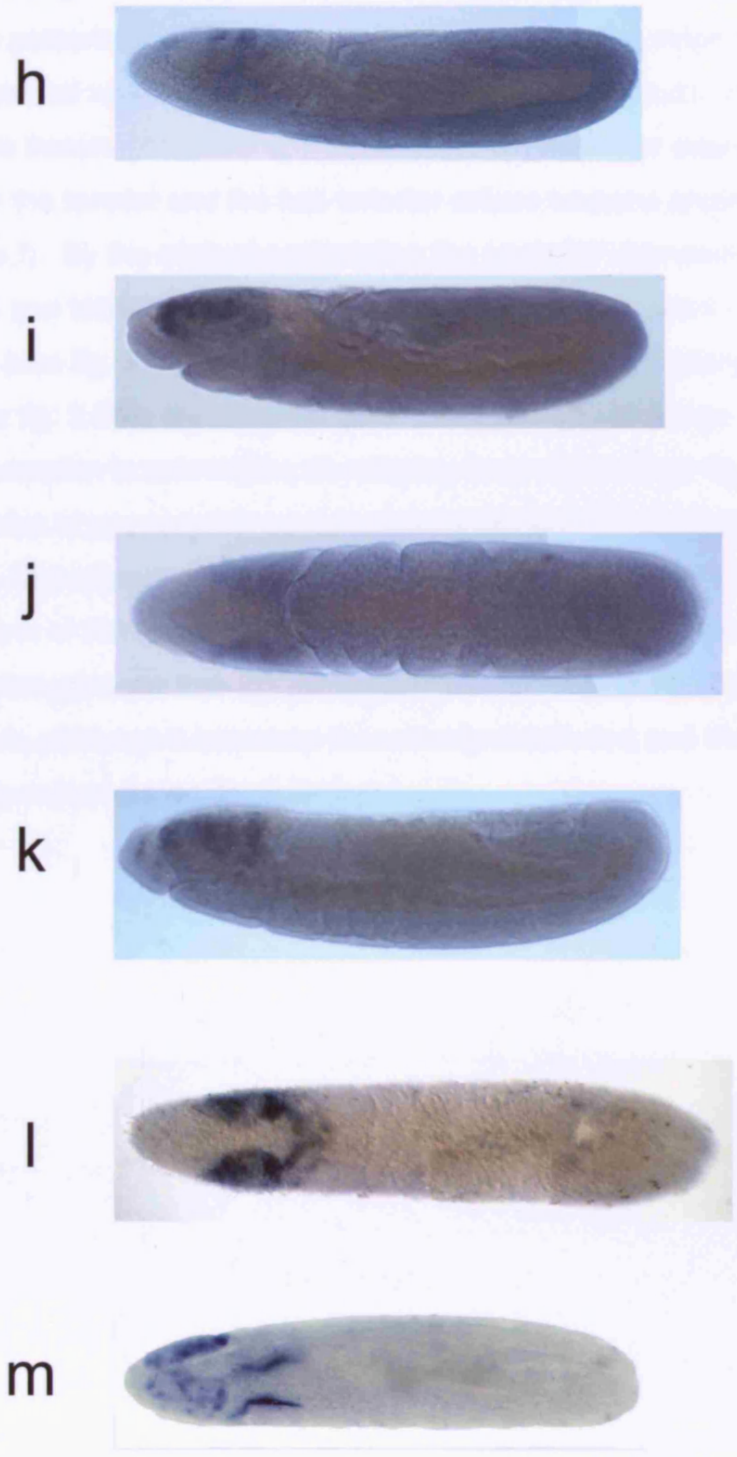


Figure 3.6
tll mRNA expression patterns in *M. domestica* embryos.
Embryos are viewed laterally, the anterior (A) to the left and posterior (P) to the right and dorsal to the top. Embryos c, f, j and l are viewed dorsally, again with anterior to the left. The different stages are as follows:
a. syncytial blastoderm b, c. & d. cellular blastoderm; e. & f. late cellular blastoderm; g. gastrulation; h, i. & j. germ band extension;
k. & l. germ band retraction; m. dorsal closure.

in the *D. melanogaster* embryos at NC 14 (see fig. 3.6b). The staining in the posterior is stronger than that in the anterior at this stage. The posterior cap then retracts posteriorly on the ventral side and the anterior stripe splits into two stripes both of which resemble horseshoes (see fig. 3.6c,d). By the end of stage 5 the transcript level drops significantly the posterior expression retracts from the termini and the two anterior stripes become angled obliquely (see fig. 3.6e,f). By the start of gastrulation the posterior expression has disappeared and faint expression of *tll* is seen in the presumptive developing brain region (see fig. 3.6g) and continues through germ band elongation (stage 8; see fig. 3.6h). By maximal germ band elongation (stage 11) stronger expression is seen in the developing brain region (see fig. 3.6i,j). This expression is uneven with certain regions being more strongly stained than others. The strongest expression is putatively in the peripheral and post cortical regions of the brain and optic lobe (as identified in *D. melanogaster)*. This expression persists through germ band retraction (see fig. 3.6k,l) and dorsal closure, although it becomes less evenly distributed and the level of expression is reduced (see fig. 3.6m).

3.3 Discussion

3.3.1 Identification of *M. domestica tll*

This chapter describes the isolation and characterization of the *M. domestica tll* gene. The putative *M. domestica* Tll protein is highly conserved with *D. melanogaster* Tll especially in the DNA binding domain. The expression patterns are also generally conserved although some differences are observed.

3.3.2 Conservation of *M. domestica* Tll

The *M. domestica* Tll protein shows conservation of the residues in the D and P boxes unique to Tll orthologues and the lysine to alanine substitution within the DNA binding domain. Indeed the DNA binding domain and T/A box are identical. This indicates that *M. domestica* Tll will bind to the same target DNA binding sequence as *D. melanogaster* Tll. The Tll ligand-binding domain is less conserved than the DNA binding domain, which is characteristic of the lower rate of conservation seen within these domains in general. However, until a ligand has been identified speculation about evolution in this domain is difficult. Other than the DNA binding domain no putative activation domains were identified in the Tll protein. Unlike some early developmental genes *tll* has not evolved long runs of amino acids, which are thought to enhance activation such as the polyglutamine tracts seen in *hb* (Bonneton *et al.*, 1997).

3.3.3 Conservation of the *tll* gene structure

There are two putative methionine start codons at the N-terminus of the *M. domestica* protein (see fig. 3.5). The *D. melanogaster* translation start was identified from three potential methionine residues due to its proximity to a sequence similar to the arthropod translation consensus (Pignoni *et al.*, 1990). The *M. domestica* sequence would confirm this choice as the first methionine since subsequent residues are conserved and a complete consensus arthropod translation sequence is found immediately upstream (see fig. 3.5).

The difficulty of deciding the translation start site by DNA sequence alone is made more pertinent by the publication of the *T. castanium tll* sequence (Schröder *et al.*, 2000). In *T. castanium tll* there are four methionine residues, each of which could potentially be the start of the protein. As usual the first in-frame ATG has been designated as the ORF start codon. However, the fourth residue is also a methionine and in an alignment coincides with the start of the *D. melanogaster* and *M. domestica* proteins (see fig. 3.2).

The non-coding regions of the *M. domestica* genes such as *hb* and *otd* are typically expanded with respect to the *D. melanogaster* sequences (Bonneton *et. al.*, 1997; McGregor, 2002). Although this is not the case for the *M. domestica tll* 5' and 3' UTRs, both the *tll* intron and the promoter regions are larger (see chapter 4). Indeed the non-coding regions in *M. domestica* are more AT rich and long tracts of A, T or AT repeats are seen. These are thought to be generated by slippage events, which may contribute to expansion of the genome and this could be related to the larger genome in *M. domestica* (Hancock *et al.*, 1999 and see 5.4.5).

3.3.4 Conservation of *tll* expression in the higher diptera

The expression patterns of *M. domestica tll* are generally similar to the expression of *D. melanogaster tll*. Fate mapping of the embryo to regions of *tll* expression in *D. melanogaster* identified the early posterior cap of expression (NC12) as being necessary for correct development of the posterior terminal region (Pignoni *et al.*, 1990). The anterior stripe at NC14 maps to those regions necessary for the correct development of the brain and the dorsal part of the cephalopharyngeal skeleton. Both these essential expression domains are conserved in *M. domestica*, which suggests the function of the *tll* gene is conserved with respect to *D. melanogaster* Tll.

The conservation of expression patterns of *tll* and other *M. domestica* genes with *D. melanogaster*, combined with the isolation of *bcd* in *M. domestica* is consistent with the conservation of *tll* regulation between these species. Therefore the other transcriptional regulators beside Bcd in *D. melanogaster*, Tor and Dl, would be predicted to be regulating *tll* in *M. domestica*. Regulation within the *M. domestica* developing brain may also be

conserved with *D. melanogaster*, but further characterization of *M. domestica* developmental genes is necessary to confirm this.

3.3.5 Differences in the *tll* expression pattern between *M. domestica* and *D. melanogaster*

Although the essential expression patterns in *D. melanogaster* are conserved in *M. domestica,* a number of differences are also observed. Consistent with changes seen between other early developmental genes, these differences appear to be in secondary expression patterns or heterochronic shifts in expression (Sommer and Tautz, 1991).

The first difference seen is that the early terminal caps of expression in *D. melanogaster tll* are missing in *M. domestica* at the syncytial blastoderm stage (see fig. 3.7a,b). It is possible that the embryos collected may have been too old to show this expression pattern, but the experiment was repeated more than once and the early *cad* expression patterns were seen (see 8.2.4).

Expression of *D. melanogaster hb* and *otd* also begin as early anterior caps of expression, which then retract from the tip (Tautz *et al.*, 1987; Finkelstein *et al.*, 1990). This expression is missing for both *M. domestica* genes (Bonneton *et al.*, 1997; McGregor, 2002). Since Tor regulates this expression of all three genes in *D. melanogaster* it is possible that derepression by Tor has become delayed in *M. domestica*. However, this delay may have no functional significance, as the expression of the anterior cap of *tll* in *D. melanogaster* is superfluous to proper development (Pignoni *et al.,* 1990). Indeed in *T. castaneum* early anterior expression of *tll* is absent, (Schröder *et al.*, 2000). However, the other primary expression patterns of *tll* are seen in *T. castaneum*; characterized by an early cap of posterior expression and later expression in the anterior most regions of the head and ocular region (Schröder *et al.*, 2000).

The anterior stripe is at first identical in *M. domestica* and *D. melanogaster* but diverges as development progresses. Although, in both species the stripes become obliquely inclined, possibly as a result of cell movement during gastrulation, in *M. domestica* the stripe becomes divided

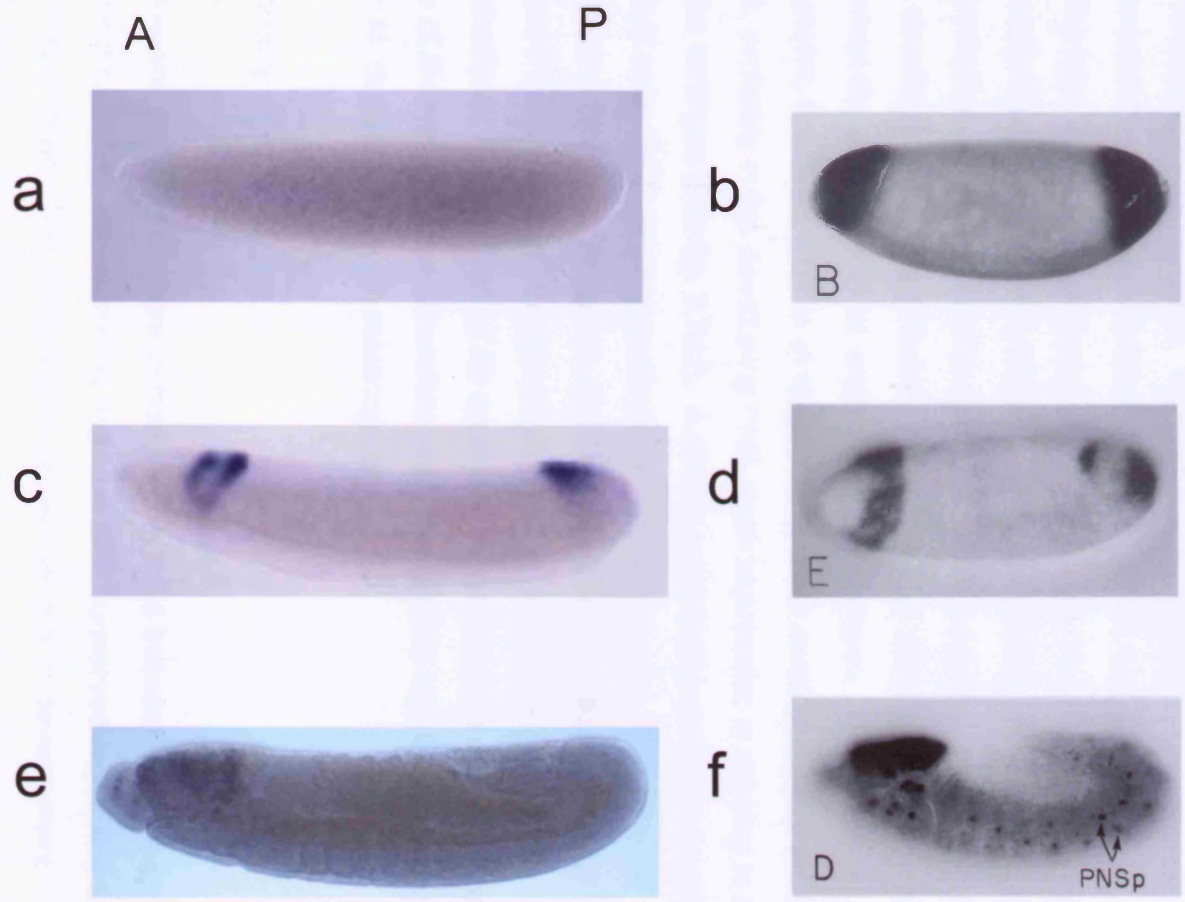


Figure 3.7
Differences in *tll* expression patterns between *M. domestica* and *D. melanogaster*.
Embryos on the left are viewed laterally, the anterior (A) to the left and posterior (P) to the right and dorsal to the top.
M. domestica and *D. melanogaster* embryos are on the left and right respectively.
The different stages are as follows:
a. and b. syncytial blastoderm;
c. and d. cellular blastoderm;
e. and f. germband retraction.
Images b, d and f are taken from Pignoni *et al,* 1990.

along the AP axis but in *D. melanogaster* along the dorsal midline (see fig. 3.7c,d). The regulator of this division in *D. melanogaster* is unknown. However, it is unlikely that the subdivisions of the stripe along different axes in the two species are regulated in the same way. The *M. domestica* pattern may result from a combination of Bicoid activation and a repressing factor. Interestingly the *M. domestica hb* gene also shows a division into a greater number of stripes than seen in *D. melanogaster*, however the functional significance of these extra stripes is unknown (Sommer and Tautz, 1991). The division of the stripe in *D. melanogaster* is likely to be a result of the subdivision of the brain into two lobes and although the stripes in *M. domestica* are not seen to split along the midline, later expression within the brain is in two separate domains. This difference in expression could be due to a heterochronic shift in the division into lobes of the developing brain.

The later expression patterns of *M. domestica tll* within the developing brain closely resemble that of *D. melanogaster tll* but with the lack of suitable tissue markers in *M. domestica* it is difficult to assign expression patterns to specific structures. This is particularly problematic in later developmental stages when the structure of the embryo is more complex. In *D. melanogaster, tll* mRNA is seen at stages 12 to 13 in small groups of cells in the trunk, probably in the developing PNS, such expression is missing in *M. domestica* embryos (see fig. 3.7e,f). A similar pattern of expression in cells of the trunk is seen for *D. melanogaster hb* mRNA and is also present in *M. domestica* and the blowfly species, *Lucilia sericata* and *Calliphora vicina* (Tautz and Pfeifle, 1989; McGregor *et al.*, 2001 and P. Shaw personal communication). Observation of *hb* expression in these cells confirms the absence of *tll* expression and therefore another possible regulatory change between *M. domestica* and *D. melanogaster*.

3.4 Summary

The expression patterns of *tll* are primarily conserved between *M. domestica* and *D. melanogaster*. The *tll* coding region is also conserved especially in the DNA binding domain. The conservation of the role of *tll* in development suggests that its position in the developmental network is conserved in the

higher dipterans. Therefore, it is likely that the regulation of *tll* is also conserved and that Bcd plays a role in this regulation. To confirm this, the regulatory regions of *tll* must be identified and then compared with those of *D. melanogaster*.

Chapter 4 Characterisation of the *tll* promoter in *M. domestica* and *D. melanogaster*

4.1 Introduction

4.1.1 Comparing the Bcd-*tll* interaction between *D. melanogaster* and *M. domestica*

A comparison of the Bcd-*hb* promoter interaction between *D. melanogaster* and *M. domestica* has revealed how a functionally conserved interaction can evolve at the sequence level (Bonneton *et al.*, 1997). Bcd regulates the expression of more than ten other genes, so it is important to consider this network of interactions when interpreting changes in the Bcd-*hb* promoter interaction. Therefore, I have expanded this study to investigate the evolution of the Bcd-*tll* promoter interaction in these two species. The Bcd-*tll* promoter interaction was chosen in part because it is more complicated than the Bcd-*hb* promoter interaction (Bonneton *et al.*, 1997; Liaw and Lengyel, 1992;). Bcd acts as both an activator and repressor of *tll* expression via separate regulatory modules (see 1.12 and fig. 1.9; Pignoni *et al.*, 1992).

4.1.2 Aims

In *D. melanogaster,* Bcd activates the anterior stripe of *tll* expression (Liaw and Lengyel, 1992). The *cis*-regulatory module responsible for the expression has been identified and consists of ten binding sites spread over approximately 100 bp (Liaw and Lengyel 1992). Isolation of the equivalent region from *M. domestica* would allow a direct comparison of the activation of *tll* by Bcd between these species.

Bcd is also involved in the repression of *tll* expression at the anterior and although the regulatory module has been identified the individual Bcd binding sites were not characterized (Liaw and Lengyel, 1992). Thus, identification of these sites would allow a comparison of the *cis*-regulatory regions that mediate activation and repression. Both types of site could also be compared between *M. domestica* and *D. melanogaster*, to identify differences between the two functions of Bcd. Therefore to make these comparisons it is necessary to identify the Bcd binding sites in the repressing region of *D. melanogaster* and both activating and repressing sites in *M. domestica*. As

the *tll* promoter is also bound by factors other than Bcd, identification of binding sites for such factors will add to the analysis of promoter function between *D. melanogaster* and *M. domestica*.

4.2 Results

4.2.1 Bcd binding sites in the *D. melanogaster tll* promoter

The region of the *D. melanogaster tll* promoter previously footprinted by Liaw and Lengyel (1992) lies between –1.3 kb and –800 bp with respect to the transcription start site. However, the sequence more proximal to the transcription start site contains the region responsive to negative regulation by Bcd (Liaw and Lengyel 1992). Therefore, to identify sites involved in Bcd repression of the *tll* promoter the region was *DNase*I footprinted with the *D. melanogaster* Bcd homeodomain (see 2.2.8).

Primers DmDNA1 to 8 were designed to cover this region, which contained 14 putative Bcd binding sites identified using the FINDPATTERNS programme (see fig. 4.1 and 2.2.12). Only four of these sites, D1 to 4, were protected in *DNase*I footprinting experiments (see fig. 4.2 and 4.3). D1 is nearest to the transcription start site (–97 bp) and D2 to D4 are positioned more closely to the Bcd activating region (see fig. 4.4A). All sites were weakly protected with only D3 and possibly D4 protected on both strands. The sequences reflect this weak protection, none of the sites matched the *D. melanogaster* consensus CTAATCC at all positions (see table 4.1), D1 is most similar with 6/7 matches and the only site with a TAAT core sequence (Driever and Nusslein-Volhard, 1989). However, the G at position 6 has never before been reported, which suggests it is unfavourable for Bcd binding (Ludwig *et al.*, 2000). Although D4 only matches the consensus at 4 positions, it is found in a head to head orientation with D3 separated by 3 bp; an arrangement suited to cooperative binding by Bcd (see 1.9). Even though sites D1-D4 have a poor match to the consensus individually, the consensus derived from all Bcd binding sites present in the *tll* promoter (D1 to D14, fig. 4.4A) is similar to that calculated for *D. melanogaster hb*. The only deviation

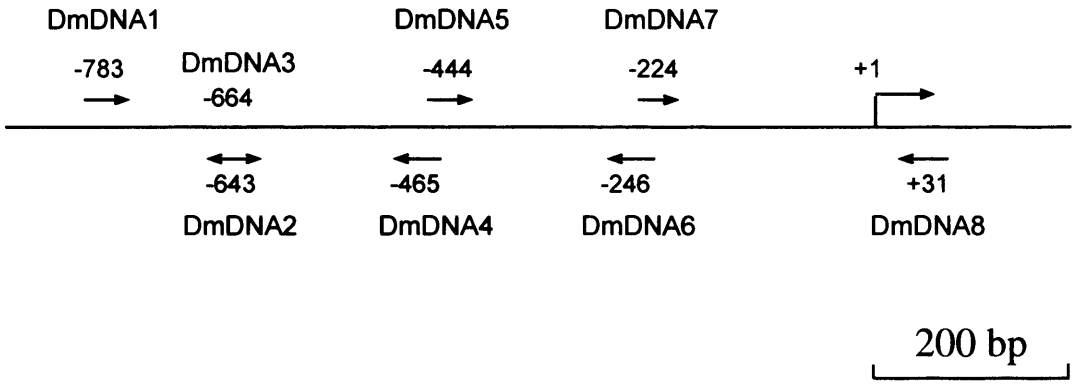


Figure 4.1 The primers used in *DNase*I footprinting of the *D. melanogaster tll* promoter. The numbering corresponds to the 5' end of the primer with respect to the transcription start, which is marked by the closed arrow. The open arrows indicate the direction of the primer in PCR reactions.

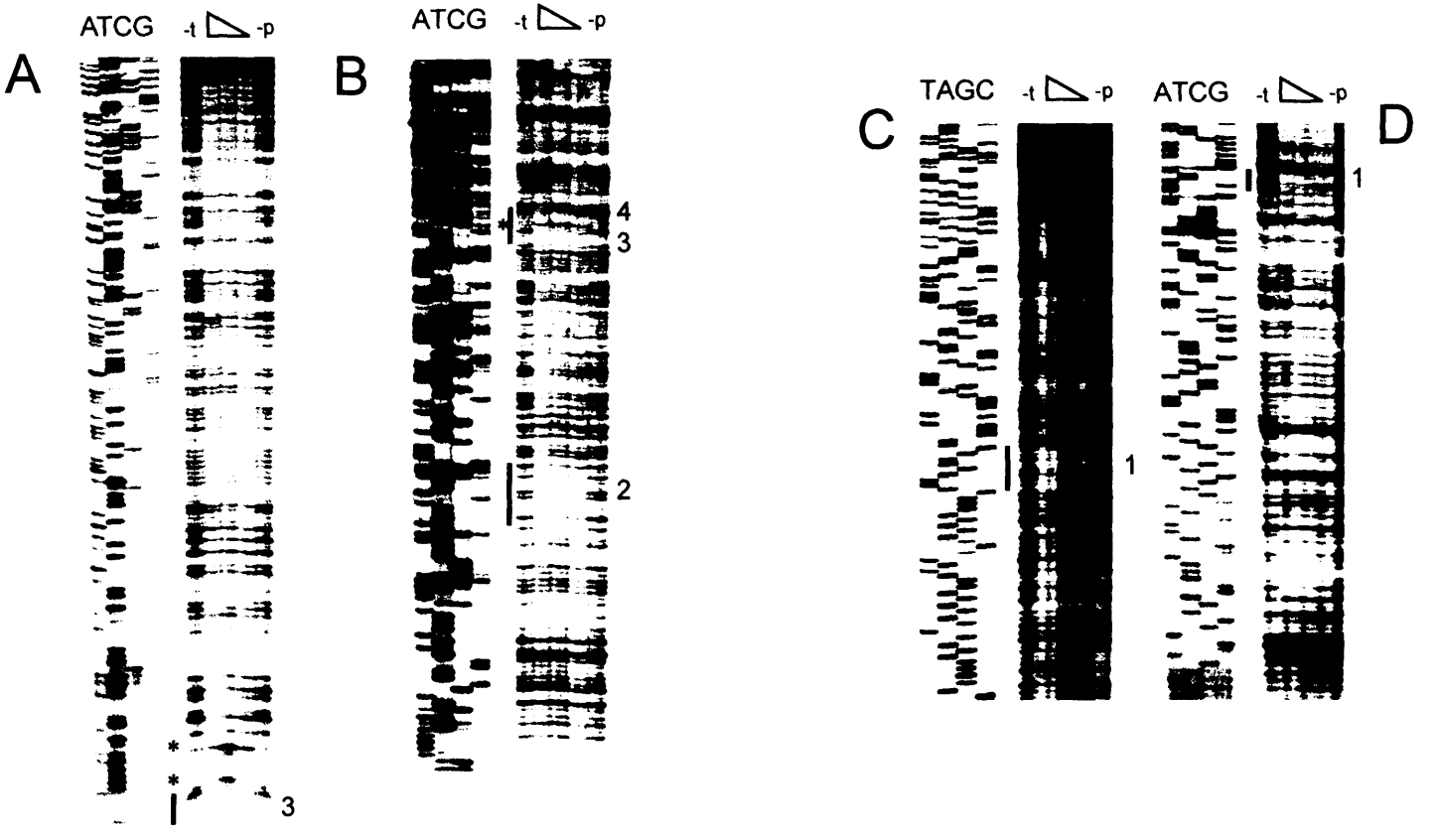


Figure 4.2 *DNaseI* Footprinting of the *D. melanogaster tll* promoter with primers DrostII3, DrostII4, DrostII7 and DrostII8 (only footprints containing Bcd sites are shown). A. Sense strand footprinting with end-labelled DrostII3. B. Antisense strand footprinting with end-labelled DrostII4. C. Sense strand footprinting with end-labelled DrostII7. D. Antisense strand footprinting with end-labelled DrostII8. DNA ladders generated by dideoxy sequencing are shown alongside. Bcd binding sites are numbered next to the footprint; black lines indicate protected regions; hypersensitive sites are shown by an asterisk. Decreasing concentrations of *D. melanogaster* Bcd homeodomain-GST protein were used in 3 reactions and are represented by the triangles above the middle 3 lanes. The control lanes are –t, GST control protein and –p, no protein added.

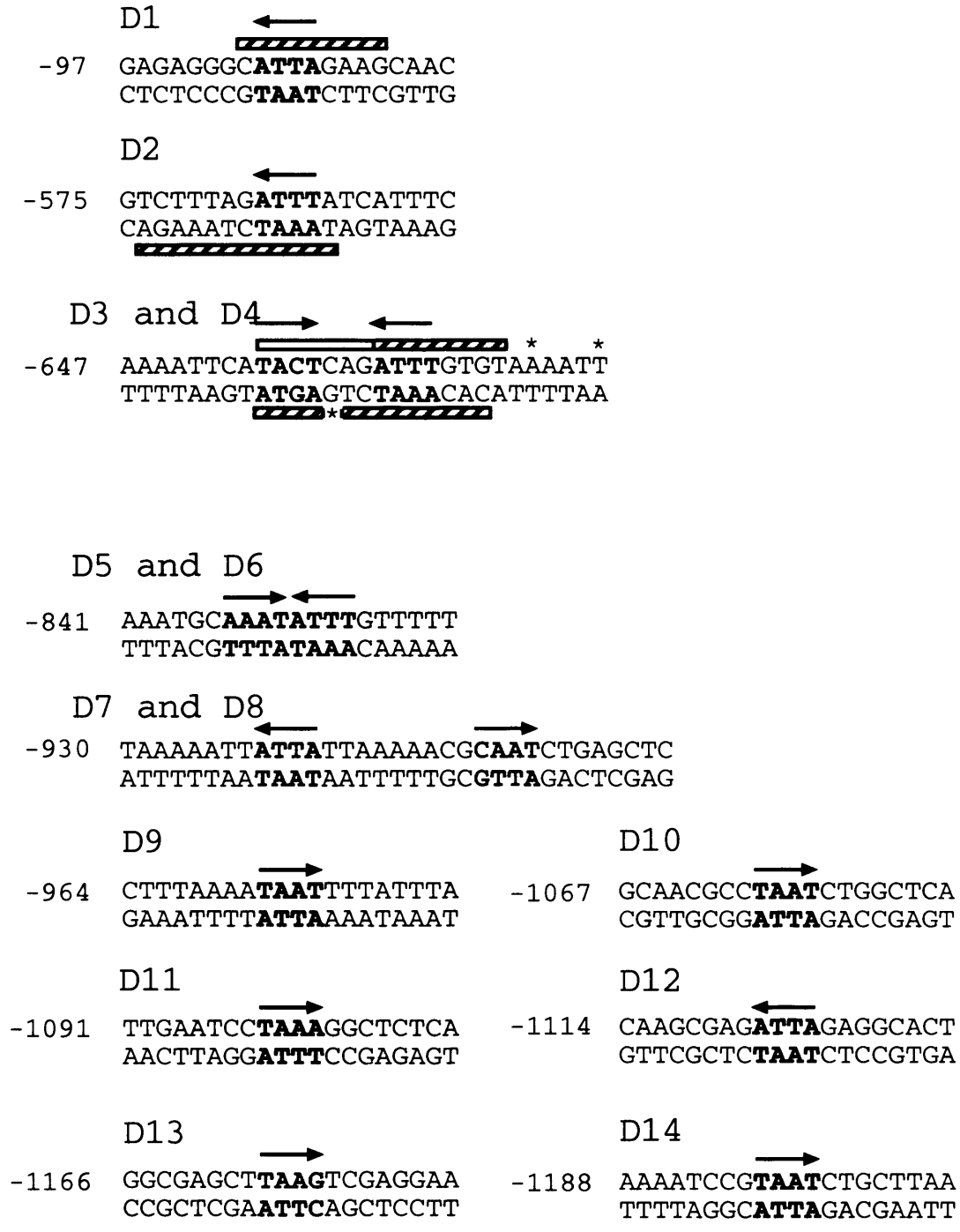


Figure 4.3 Sequences of the Bcd binding sites in the *D. melanogaster tll* promoter. Sites D1 to D4 were found in this study, the other sites correspond to sites 1 to 8 in Liaw and Lengyel 1992. Binding sites are numbered with respect to the text (see table 4.1) and their position is given with respect to the transcription start site and corresponds to the first base of the core sequence (in bold). Where two sites are shown together the numbering refers to the upstream site. The core sequences are shown in bold, black for a TAAT core and blue for a non-TAAT core. Arrows above the sequences indicate the orientation of the binding site. Hashed boxes indicate the regions protected in *DNaseI* experiments and hypersensitive sequences (*) are shown. The white boxes represent regions at which the state of protection was ambiguous.

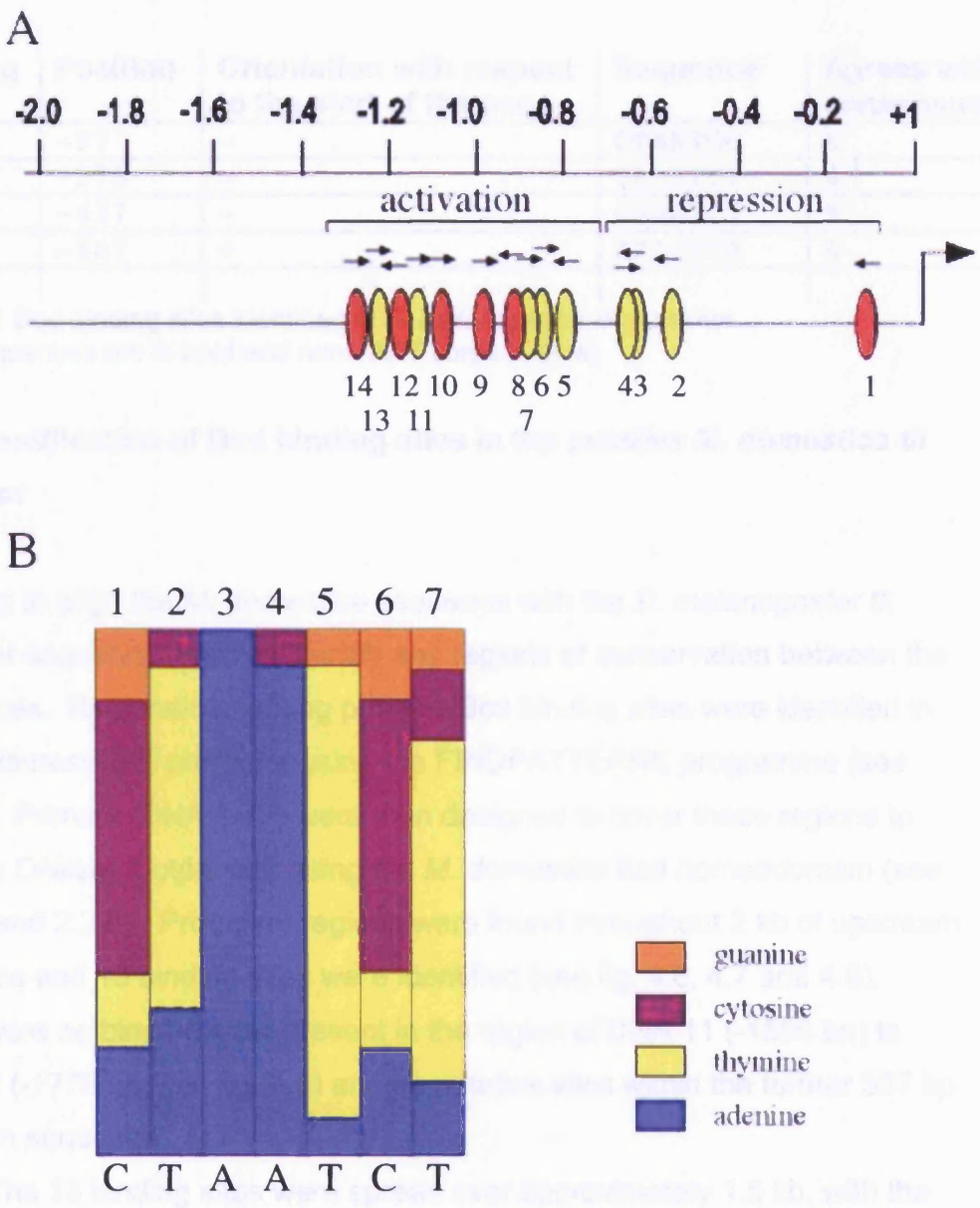


Figure 4.4A. The arrangement of Bcd binding sites in the *D. melanogaster tll* promoter The Bcd binding sites discovered in this work are numbered 1 to 4. The other sites correspond to sites 1 to 8 in Liaw and Lengyel 1992. The ovals represent Bcd binding sites, red for a TAAT core and yellow for a non-TAAT core, the small arrows indicate the orientation of the site. The large arrow indicates the transcription start site and the scale is in kilobases.
B. Calculation of the *D. melanogaster tll* Bcd binding site sequences. This figure shows the fraction of each of the 4 bases present at each of the 7 positions of the Bcd binding site sequence. The numbering above refers to the bases of the Bcd binding site sequence and below is shown the consensus sequence. The coloured boxes represent the fraction of each of the four bases present at that site, calculated from all Bcd binding sites in the *D. melanogaster tll* promoter; adenine in blue, thymine in yellow, cytosine in purple and guanine in orange.

is a T at position 7 as opposed to a C (Driever and Nusslein-Volhard, 1988; see fig. 4.4B).

Binding site	Position	Orientation with respect to the start of the gene	Sequence	Agrees with consensus ($^{x}/_{7}$)
D1	-97	-	**CTAAT**GC	6
D2	-575	-	**TAAATCT**	4
D3	-637	-	**CAAATCT**	5
D4	-647	+	**ATACTCA**	4

Table 4.1 Bcd binding sites identified in *D. melanogaster tll* promoter (core sequences are in bold and non TAAT cores in blue)

4.2.2 Identification of Bcd binding sites in the putative *M. domestica tll* promoter

Attempts to align the *M. domestica* sequence with the *D. melanogaster tll* promoter sequence failed to identify any regions of conservation between the sequences. Regions containing putative Bcd binding sites were identified in the *M. domestica tll* promoter using the FINDPATTERNS programme (see 2.2.12). Primers (DNA 1-12) were then designed to cover these regions to facilitate *DNase*I footprinting using the *M. domestica* Bcd homeodomain (see fig. 4.5 and 2.2.8). Protected regions were found throughout 2 kb of upstream sequence and 13 binding sites were identified (see fig. 4.6, 4.7 and 4.8). There were no binding sites present in the region of DNA 11 (-1566 bp) to DNA 12 (-1778 bp; see fig. 4.5) and no putative sites within the further 337 bp of known sequence.

The 13 binding sites were spread over approximately 1.5 kb, with the first site positioned at –50 bp with respect to the transcription start site (see fig. 4.9). The binding sites were arranged in four clusters (see fig. 4.9). Binding sites were considered to be in a cluster if they were separated by 100 bp or less. This is because Bcd is unable to bind cooperatively to binding sites spaced further than 100 bp apart (Ma *et al.*, 1996; see fig. 4.10). The sites were orientated in both directions with respect to the transcription start site.

Of the 13 sites only two (7 and 8) matched the *M. domestica* Bcd binding site consensus TTAATCY completely, but another three matched at

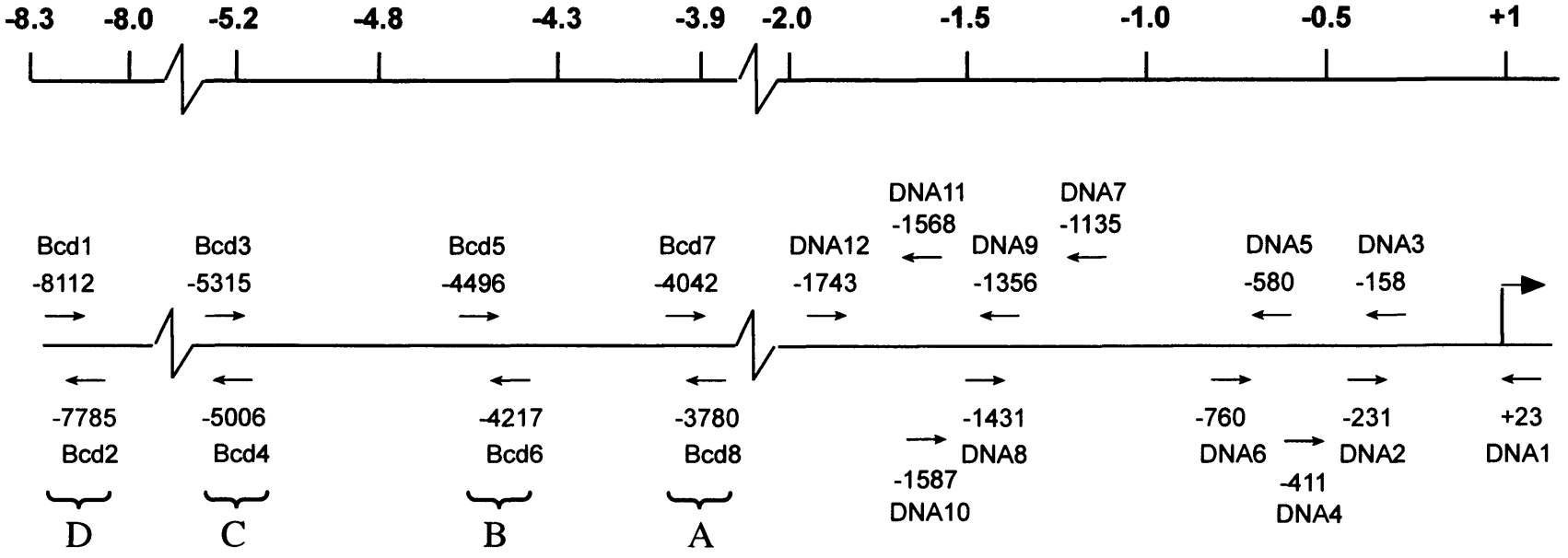


Figure 4.5 The primers used in *DNaseI* footprinting of the *M. domestica tll* promoter. The numbering corresponds to the 5' end of the primer with respect to the transcription start, which is marked by the closed arrow. The open arrows indicate the direction of the primer in PCR reactions. Footprinted Regions A-D are marked.

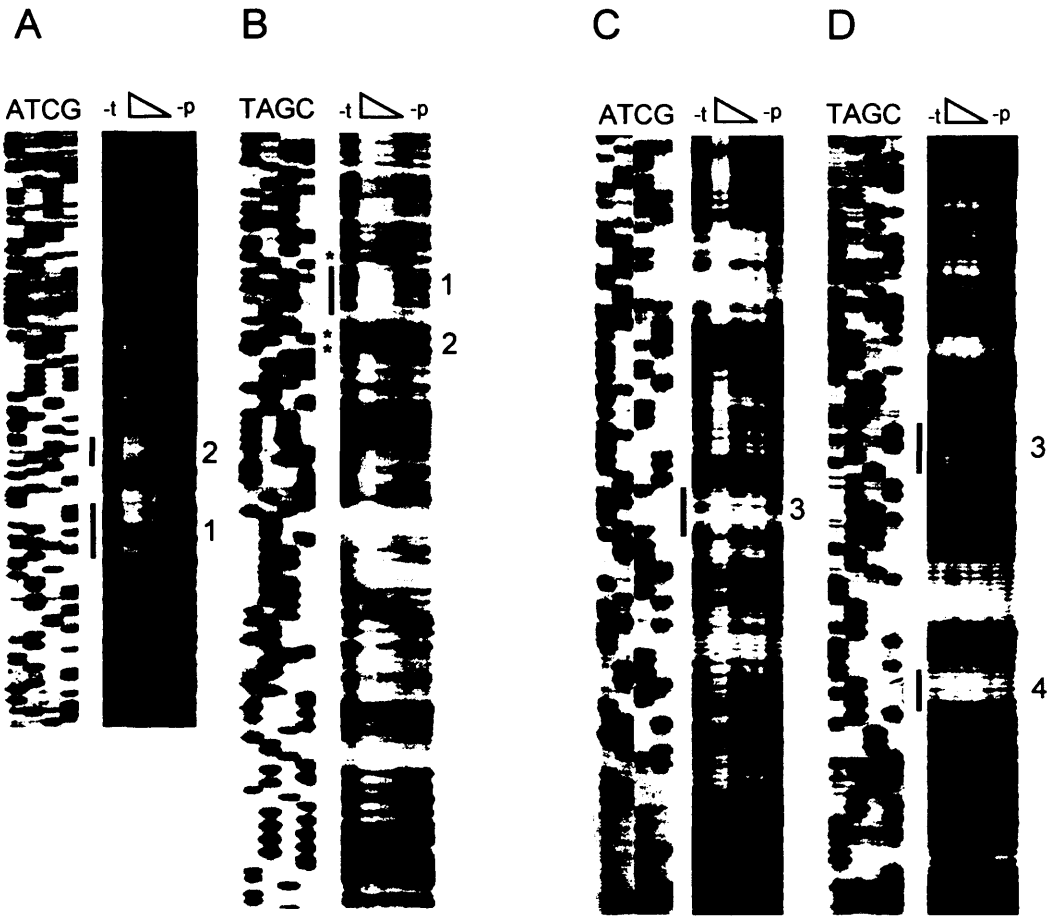


Figure 4.6 *DNaseI* Footprinting of the *M. domestica tll* promoter with primers DNA1, DNA2, DNA3 and DNA4 (only footprints containing Bcd sites are shown). A. Antisense strand footprinting with end-labelled DNA1. B. Sense strand footprinting with end-labelled DNA2. C. Antisense strand footprinting with end-labelled DNA3. D. Sense strand footprinting with end-labelled DNA4.
DNA ladders generated by dideoxy sequencing are shown alongside. Bcd binding sites are numbered next to the footprint; black lines indicate protected regions; hypersensitive sites are shown by an asterisk. Decreasing concentrations of *M. domestica* Bcd homeodomain-GST were used in 3 reactions and are represented by the triangles above the middle 3 lanes. The control lanes are –t, GST control protein and –p, no protein added.

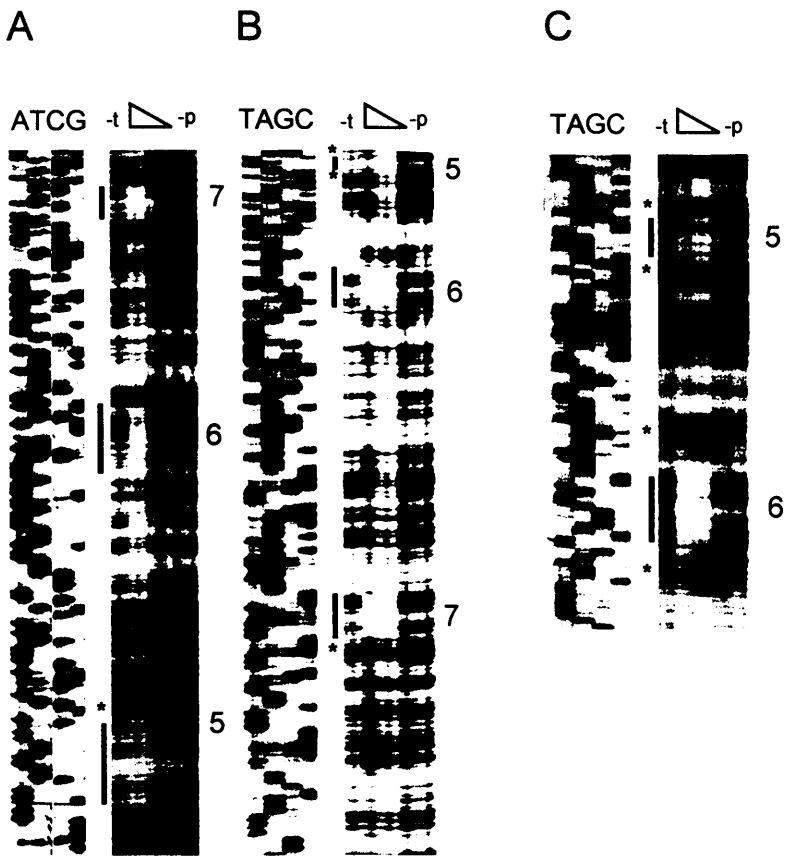


Figure 4.7 *DNase*I Footprinting of the *M. domestica tll* promoter with primers DNA5 and DNA6 (only footprints containing Bcd sites are shown). A. Antisense strand footprinting with end-labelled DNA5. B. & C. Sense strand footprinting with end-labelled DNA6. In B. the gel was run for longer to get a greater resolution.
DNA ladders generated by dideoxy sequencing are shown alongside. Bcd binding sites are numbered next to the footprint; black lines indicate protected regions; hypersensitive sites are shown by an asterisk. Decreasing concentrations of *M. domestica* Bcd homeodomain-GST were used in 3 reactions and are represented by the triangles above the middle 3 lanes. The control lanes are –t, GST control protein and –p, no protein added.

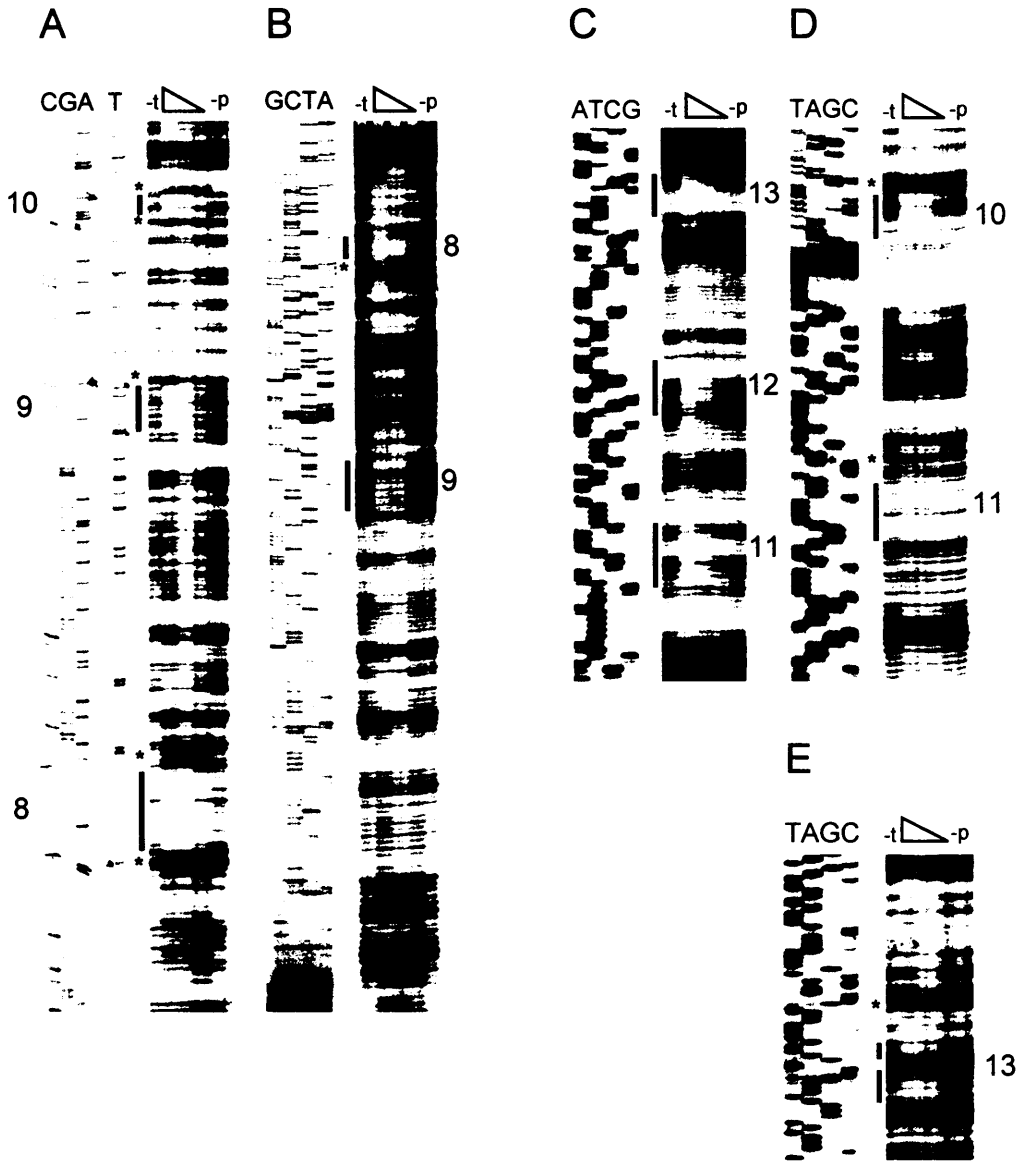


Figure 4.8 *DNase*I Footprinting of the *M. domestica tll* promoter with primers DNA7, DNA8, DNA9 and DNA10 (only footprints containing Bcd sites are shown). A. Antisense strand footprinting with end-labelled DNA7. B. Sense strand footprinting with end-labelled DNA8. C. Antisense strand footprinting with end-labelled DNA9. D. and E. Sense strand footprinting with end-labelled DNA10. In D. the gel was run for longer to get a greater resolution. DNA ladders generated by dideoxy sequencing are shown alongside. Bcd binding sites are numbered next to the footprint; black lines indicate protected regions; hypersensitive sites are shown by an asterisk. Decreasing concentrations of *M. domestica* Bcd homeodomain-GST were used in 3 reactions and are represented by the triangles above the middle 3 lanes. The control lanes are –t, GST control protein and –p, no protein added.

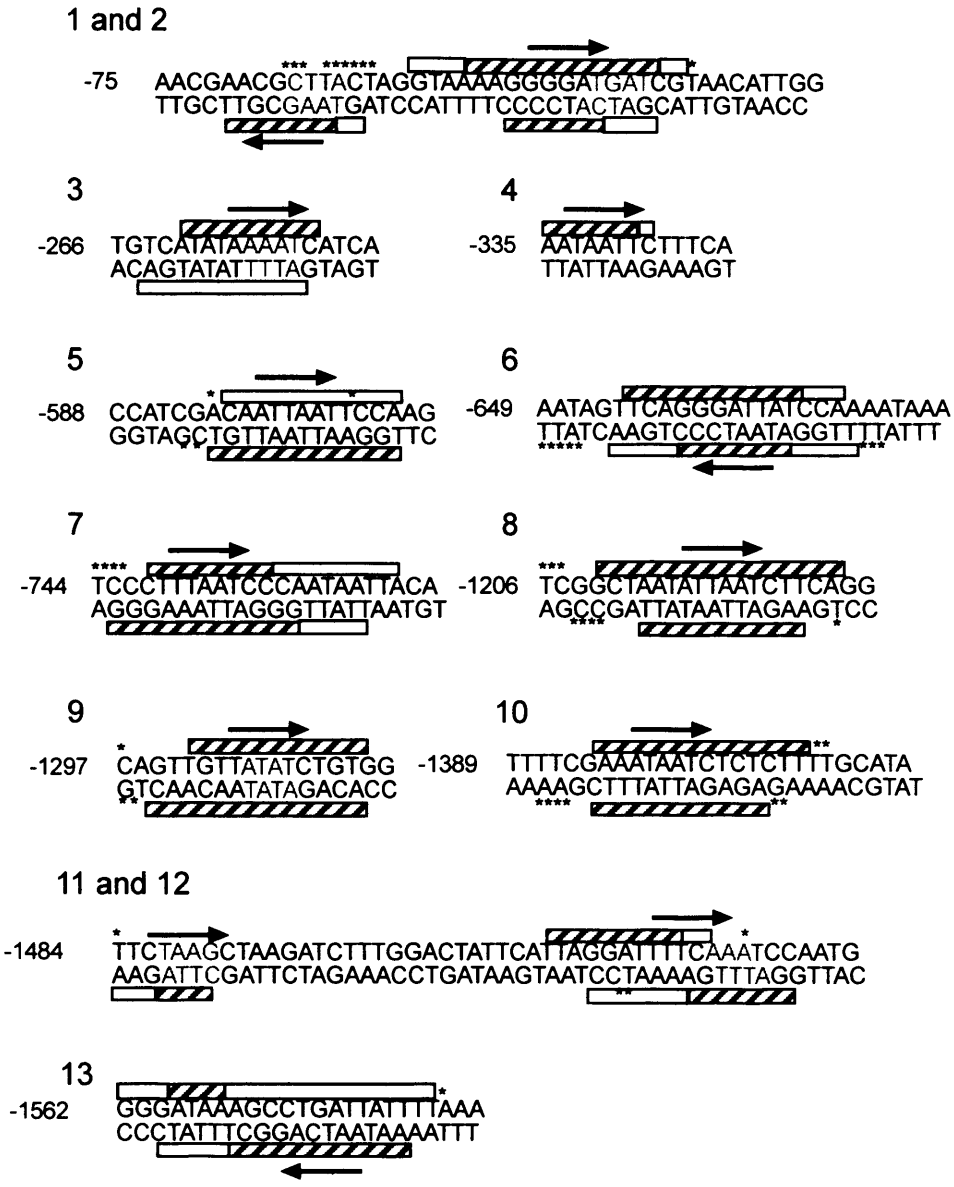


Figure 4.9 Sequences of the Bcd binding sites in the *M. domestica tll* promoter. Binding sites are numbered with respect to the text (see table 4.2) and their position is given with respect to the transcription start site and corresponds to the first base of the core sequence (in bold). Where two sites are shown together the numbering refers to the upstream site. The core sequences are shown in bold, black for a TAAT core and blue for a non-TAAT core. Arrows above the sequences indicate the orientation of the binding site. Hashed boxes indicate the regions protected in *DNase*I experiments and hypersensitive sequences (*) are shown. The white boxes represent regions at which the state of protection was ambiguous.

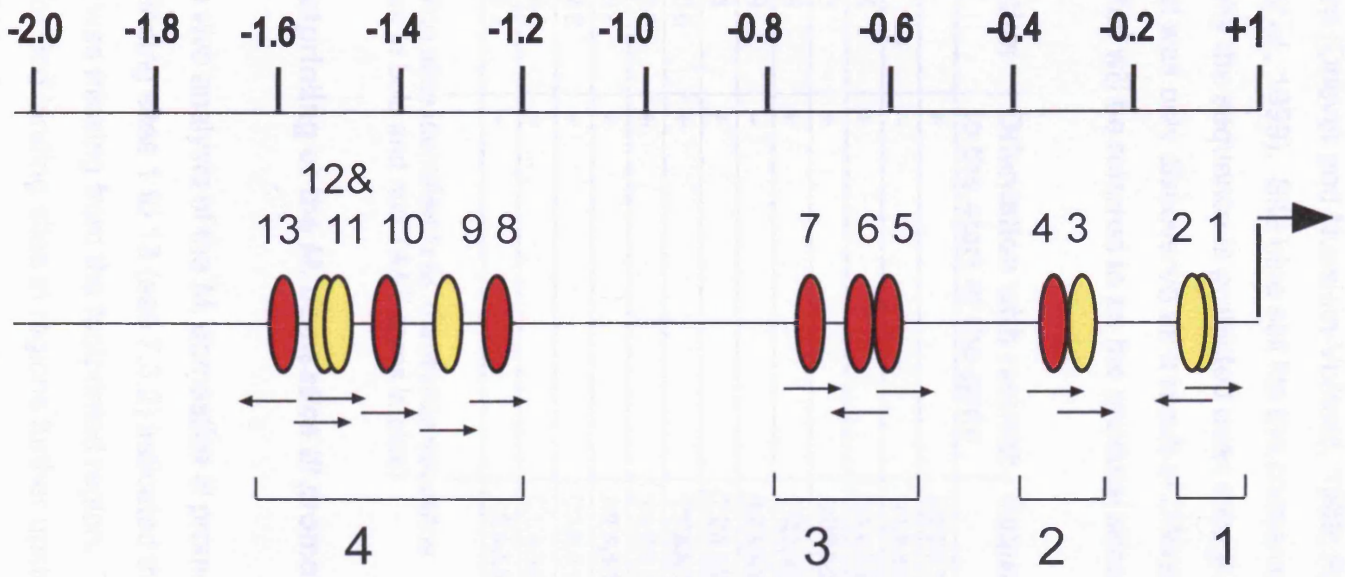


Figure 4.10 The arrangement of Bcd binding sites in the *M. domestica tll* promoter. The ovals represent Bcd binding sites, red for a TAAT core and yellow for a non-TAAT core, the small arrows indicate the orientation of the site. The large arrow indicates the transcription start site and the scale is in kilobases.

6/7 bases (Bonneton *et al.*, 1997; see table 4.2). Seven out of 13 of the sites contained the core TAAT sequence. Of the different core sequences, TGAT, TAAG and AAAT have been found before but site nine was unusual by having a novel ATAT core (Driever and Nusslein-Volhard, 1989; Rivera-Pomar *et al.*, 1995 and Yuan *et al.*, 1999). Site nine still fits the consensus at 5/7 positions, which explains why the sequence is protected even though it was not a predicted site and was only discovered as a result of *DNase*I footprinting. These thirteen sites will be referred to as the proximal sites.

Binding site	Position	Orientation with respect to the start of the gene	Sequence	Agrees with consensus ($^{x}/_{7}$)
1	-59	+	ATGATCG	4
2	-75	-	GTAAGCG	4
3	-266	+	AAAATCA	4
4	-335	+	ATAATTC	5
5	-588	+	TTAATTC	6
6	-649	-	ATAATCC	6
7	-744	+	TTAATCC	7
8	-1206	+	TTAATCT	7
9	-1297	+	TATATCT	5
10	-1389	+	ATAATCT	6
11	-1448	+	CAAATCC	5
12	-1484	+	CTAAGCT	5
13	-1562	-	ATAATCA	5

Table 4.2 Bcd binding sites identified in *M. domestica* sequence (core sequences are in bold and non TAAT cores in blue)

4.2.3 Further footprinting of the *M. domestica tll* promoter

Evidence from *in vivo* analysis of the *M. domestica tll* promoter sequences containing Bcd binding sites 1 to 13 (see 7.3.2) indicated that part of the Bcd activating region was missing from the footprinted region. Therefore, I decided to look for Bcd binding sites in regions further upstream.

A further 7 kb of sequence was analysed for putative Bcd binding sites. Some of the putative sites were found in clusters whilst others were far from adjacent sites. The *D. melanogaster* activating region contains a cluster of Bcd binding sites. Therefore, four regions of the *M. domestica tll* upstream sequences containing clusters of predicted sites were chosen to be footprinted (see fig. 4.5, regions A to D). Primers Bcd1 to 8 were designed

(see fig. 4.5) and the corresponding regions footprinted (see fig. 4.11, 4.12 and 4.13). All four regions contained Bcd binding sites and 17 sequences were identified most of which were protected on both strands (see fig. 4.14). These sites will be referred to as the distal sites to distinguish them from the 13 sites already known (proximal sites). There are up to a maximum of 20 distal binding sites present because at three sites both strands were protected and there was a Bcd recognition sequence on each strand (see sites 14/15, 22/23, and 24/25, fig. 4.14).

Regions C and D contained three sites each arranged as a pair with the third site more distantly spaced (see fig. 4.14 and 4.15). Region A contained three or four sites spread over 50 bp. Region B was the most remarkable containing eight to ten sites within 190 bp (see fig. 4.15). The density of Bcd protein bound to region B in the footprinting experiment resulted in large footprints, with almost continuous protection over 70 bp on one strand (see fig. 4.11).

The sites in B are arranged in a manner that is highly favourable for cooperative binding. There are three pairs of sites that lie tail-to-tail separated by 19, 15 and 13 bp (see fig. 4.15). However this is the region in which three pairs of overlapping Bcd recognition sequences are present. Therefore two different interpretations could be a tail-to-tail arrangement of two pairs separated by 19 and 13 bp and a head-to-head pair of either 2 or 5 bp (see fig. 4.15). It is possible that all of these arrangements are functional depending on which site is bound first. All arrangements should result in a strong affinity for the binding sites in this region and this suggests that this region is functionally equivalent to the activating region in the *D. melanogaster tll* promoter (Yuan *et al.*, 1999).

Of the 20 distal sites only two fit the *M. domestica* Bcd consensus sequence at all sites (Bonneton *et al.,* 1997; see table 4.3). Nine of the sites have 6/7 matches and a further seven sites had 5/7 matches. As observed with the proximal binding site sequences, many of the sites did not have a perfect TAAT core (see table 4.3, blue core sequences). An equivalent number of sites contained the common alternative TAAG and there were also three sites with an AAAT core. Of all reported Bcd binding sites in *D. melanogaster* none has a CAAT core but one was observed in *M. domestica*

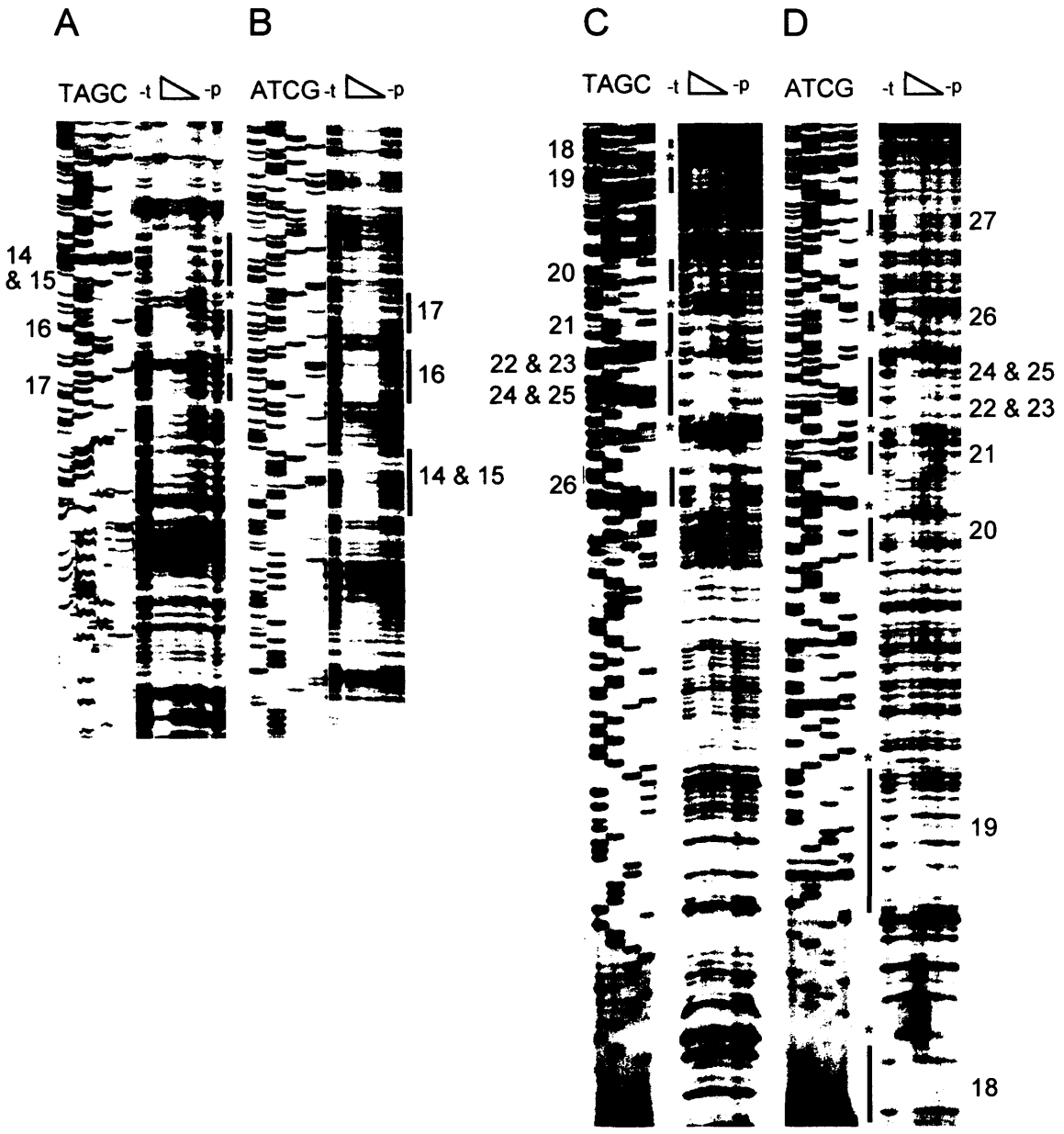


Figure 4.11 *DNase*I Footprinting of the *M. domestica tll* promoter region A with primers Bcd7 and Bcd8 and region B with primers Bcd5 and Bcd6 (only footprints containing Bcd sites are shown). A. Sense strand footprinting with end-labelled Bcd7. B. Antisense strand footprinting with end-labelled Bcd8. C. Sense strand footprinting with end-labelled Bcd5. D. Antisense strand footprinting with end-labelled Bcd6.
DNA ladders generated by dideoxy sequencing are shown alongside. Bcd binding sites are numbered next to the footprint; black lines indicate protected regions; hypersensitive sites are shown by an asterisk. Decreasing concentrations of *M. domestica* Bcd homeodomain-GST were used in 3 reactions and are represented by the triangles above the middle 3 lanes. The control lanes are –t, GST control protein and –p, no protein added.

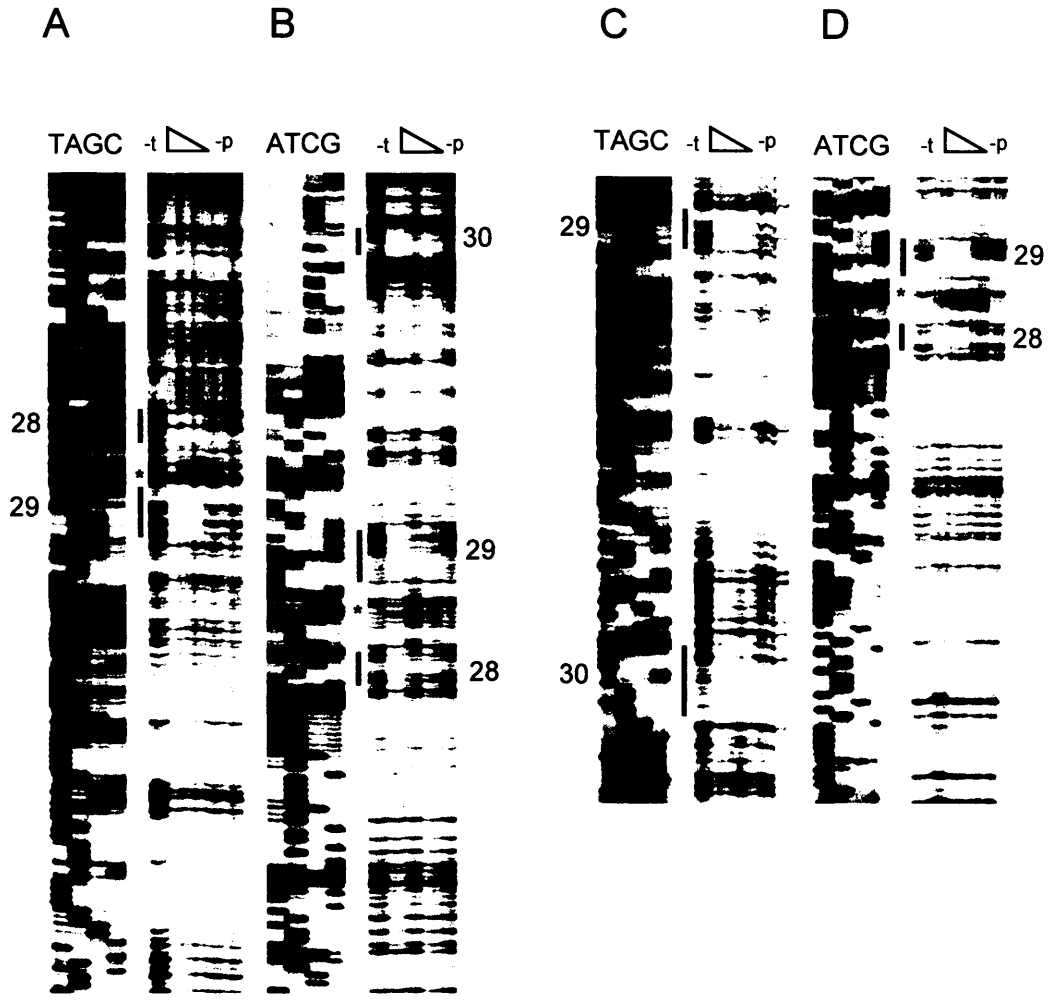


Figure 4.12 *DNase*I Footprinting of the *M. domestica tll* promoter region C with primers Bcd3 and Bcd4 (only footprints containing Bcd sites are shown). A. & C. Sense strand footprinting with end-labelled Bcd3. B. & D. Antisense strand footprinting with end-labelled Bcd4. In C. & D. the gels were ran for longer to get a greater resolution.
DNA ladders generated by dideoxy sequencing are shown alongside. Bcd binding sites are numbered next to the footprint; black lines indicate protected regions, hypersensitive sites are shown by an asterisk. Decreasing concentrations of *M. domestica* Bcd homeodomain-GST were used in 3 reactions and are represented by the triangles above the middle 3 lanes. The control lanes are –t, GST control protein and –p, no protein added.

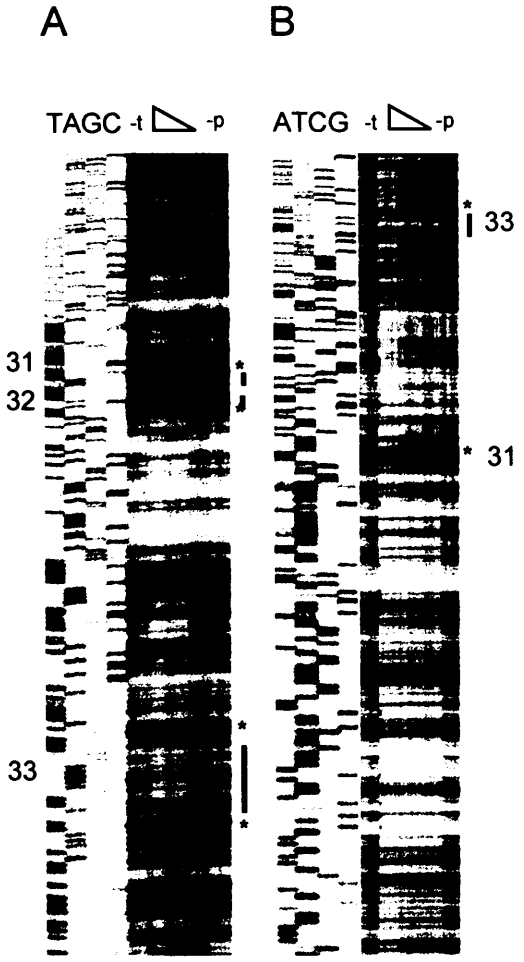


Figure 4.13 *DNase*I Footprinting of the *M. domestica tll* promoter region D with primers Bcd1 and Bcd2 (only footprints containing Bcd sites are shown). A. Sense strand footprinting with end-labelled Bcd1. B. Antisense strand footprinting with end-labelled Bcd2.
DNA ladders generated by dideoxy sequencing are shown alongside. Bcd binding sites are numbered next to the footprint; black lines indicate protected regions; hypersensitive sites are shown by an asterisk. Decreasing concentrations of *M. domestica* Bcd homeodomain-GST were used in 3 reactions and are represented by the triangles above the middle 3 lanes. The control lanes are –t, GST control protein and –p, no protein added.

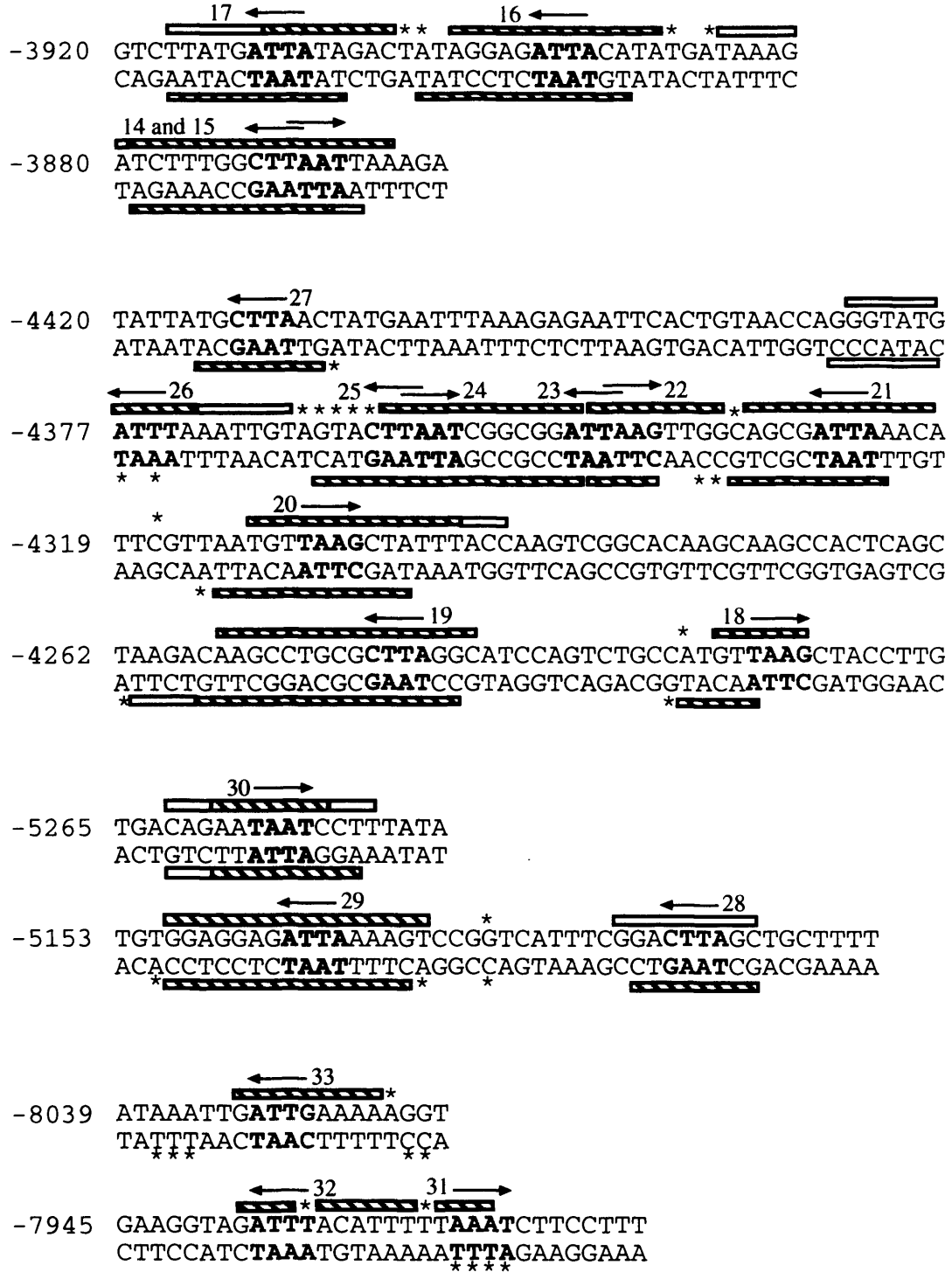


Figure 4.14 Sequences of the Bcd binding sites in regions A-D of the *M. domestica tll* promoter. Binding sites are numbered with respect to the text (see table 4.3) and their position is given with respect to the transcription start site and corresponds to the first base of the core sequence (in bold). Where two sites are shown together the numbering refers to the upstream site. The core sequences are shown in bold, black for a TAAT core and blue for a non-TAAT core. Arrows above the sequences indicate the orientation of the binding site. Hashed boxes indicate the regions protected in *DNase*I experiments and hypersensitive sequences (*) are shown. The white boxes represent regions at which the state of protection was ambiguous.

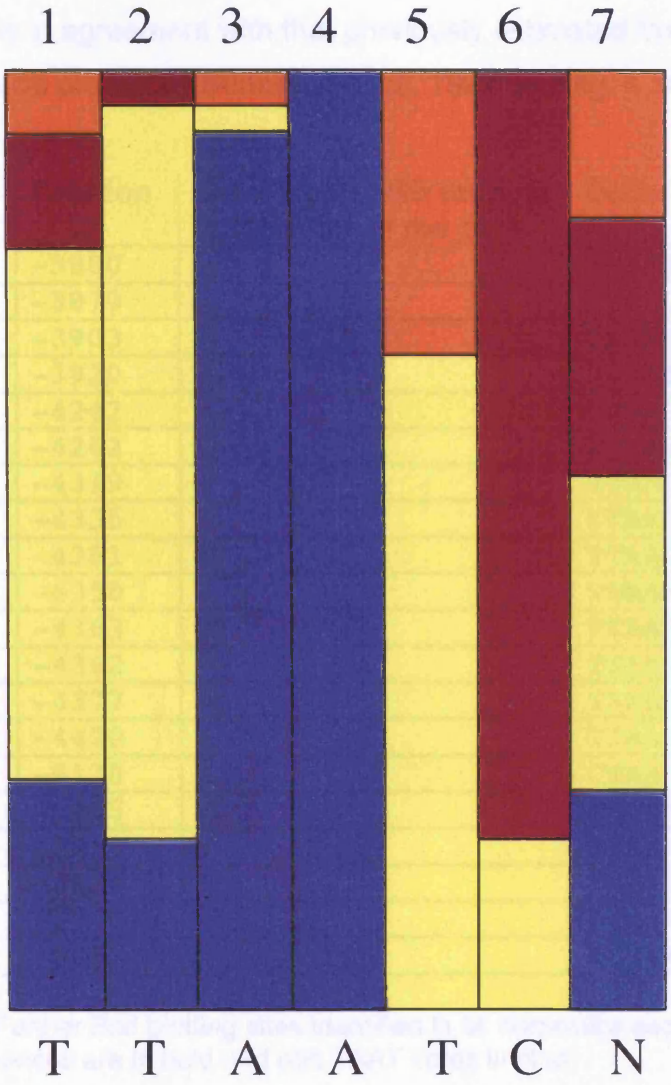


Figure 4.15 Calculation of the *M. domestica tll* Bcd binding site sequences. This figure shows the fraction of each of the 4 bases present at each of the 7 positions of the Bcd binding site sequence. The numbering above refers to the bases of the Bcd binding site sequence and below is shown the consensus sequence. The coloured boxes represent the fraction of each of the four bases present at that site, calculated from all Bcd binding sites in the *M. domestica tll* promoter; adenine in blue, thymine in yellow, cytosine in purple and guanine in orange.

region D although this was protected on only one strand (Ludwig *et al.*, 2000). The consensus sequence calculated for all Bcd sites in the *M. domestica tll* promoter is in agreement with that previously estimated from sites in the *M. domestica hb* promoter (Bonneton *et al.*, 1997; see fig. 4.16).

Binding site	Position	Orientation with respect to the start of the gene	Sequence	Agrees with consensus ($^{x}/_{7}$)
14	-3880	+	**TTAAT**TA	5
15	-3879	-	TTAAGCC	6
16	-3903	-	G**TAAT**CT	6
17	-3920	-	A**TAAT**CA	5
18	-4242	+	TTAAGCT	6
19	-4262	-	CTAAGCG	4
20	-4319	+	TTAAGCT	6
21	-4335	-	T**TAAT**CG	6
22	-4351	+	TTAAGTT	5
23	-4350	-	T**TAAT**CC	7
24	-4363	+	T**TAAT**CG	6
25	-4362	-	TTAAGTA	5
26	-4377	-	TAAATCA	5
27	-4420	-	TTAAGCA	5
28	-5130	-	CTAAGTC	4
29	-5153	-	T**TAAT**CT	7
30	-5265	+	A**TAAT**CC	6
31	-7936	+	TAAATCT	6
32	-7945	-	TAAATCT	6
33	-8039	-	TCAATCA	5

Table 4.3 Further Bcd binding sites identified in *M. domestica* sequence (core sequences are in bold and non TAAT cores in blue)

A complete comparison of all Bcd binding sites identified in *D. melanogaster* and *M. domestica* is shown in fig. 4.17. The *D. melanogaster tll* 1.6 kb Bcd activating region does not extend further than approximately 2.3 kb from the transcription start. The core-activating region from –0.8 kb to –1.2 kb contains nine closely spaced sites (Liaw and Lengyel, 1992). The putative *M. domestica* Bcd activating region extends over 3 kb further from the transcription start site than in *D. melanogaster*. The putative core group consists of eleven closely spaced sites in 551 bp (see fig. 4.17).

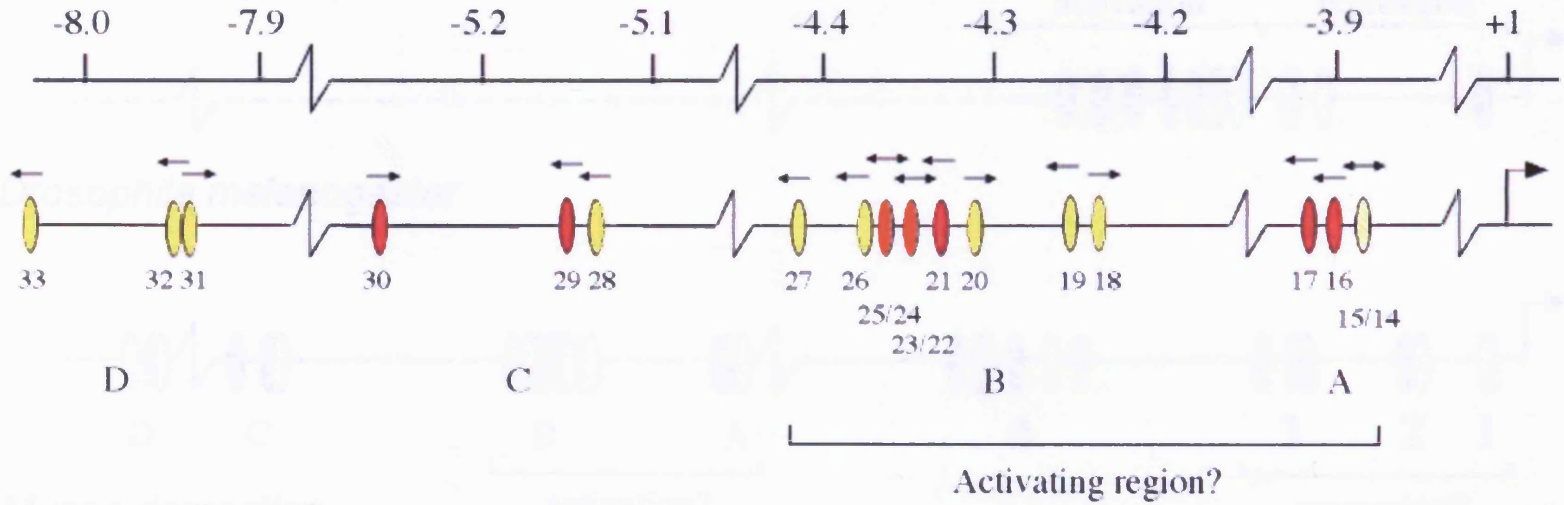


Figure 4.16 A comparison of Bcd binding sites in the *D. melanogaster* and *M. domestica tll* promoters. The regions A-D of the *M. domestica tll* promoter are marked. Red ovals correspond to TAAT core sites and yellow ovals to non-TAAT core sites, small arrows indicate the orientation of the binding sites. The arrow indicates the transcription start site. The regions not footprinted are not shown and are indicated by " -^\/- " and the scale is in kilobases

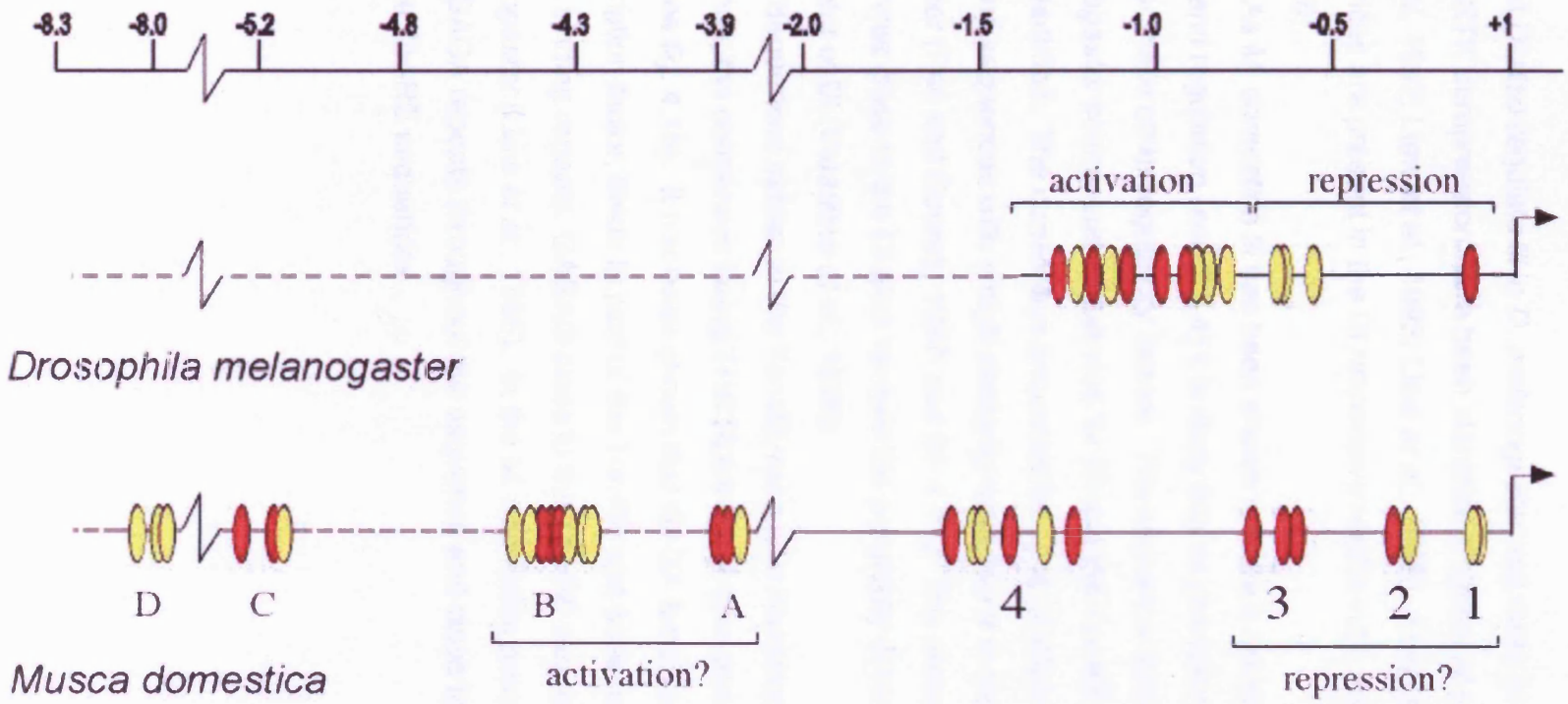


Figure 4.17 A comparison of Bcd binding sites in the *D. melanogaster* and *M. domestica tll* promoters. The regions 1-4 and A-D of the *M. domestica tll* promoter are marked. Red ovals correspond to TAAT core sites and yellow ovals to non-TAAT core sites, small arrows indicate the orientation of the binding sites. The arrow indicates the transcription start site. Putative or known activating and repressing regions are shown. The regions not footprinted are shown by the dotted lines, this includes the sequences not shown (-∿-) and the scale is in kilobases.

4.2.4 Evidence for the regulation of *tll* by Tor and Dl.

Tor and Dl also regulate *tll* in *D. melanogaster* and three elements responsive to Tor-RTK derepression have been identified (Pignoni *et al.*, 1992; Liaw and Lengyel, 1992; Liaw *et al.*, 1993; Liaw *et al.*, 1995). Two Dl consensus sequences are present in the Dl responsive region in *D. melanogaster* (see fig. 4.18).

As *M. domestica tll* has been shown to have a conserved function (see 3.3.4) and regulation (see 7.3.4) it is likely that its promoter contains binding sites for these other regulatory factors. The sequence was analysed for *D. melanogaster* consensus sequences for Dl and the Tor-RE and a number of sites identified. The consensus sequence for Dl is GGG(A/T)$_n$CC(A/C), (n = 4 or 5) and sequences with a high similarity were found in the *M. domestica tll* promoter (Pan and Courey, 1992; see fig. 4.18). The presence of AT rich sequences close to the Dl sites reveals the possibility of binding by Dri a co-repressor of Dl (Valentine *et al.*, 1998).

Sequences similar to the Tor-RE were also identified in the promoter sequence, the consensus being TG**CTCAAT**GAA (the core sequence is in bold; see fig. 4.18). It has been shown that GAGA factor, a general transcription factor, binds to part of the Tor-RE and there are also many GAGA binding repeats, GAGAG close to the Tor-RE sequences in *D. melanogaster* (Liaw *et al.*, 1995). In the *M. domestica* promoter there are many GAGA repeats throughout the sequence and close to some of the putative Tor-RE sequences.

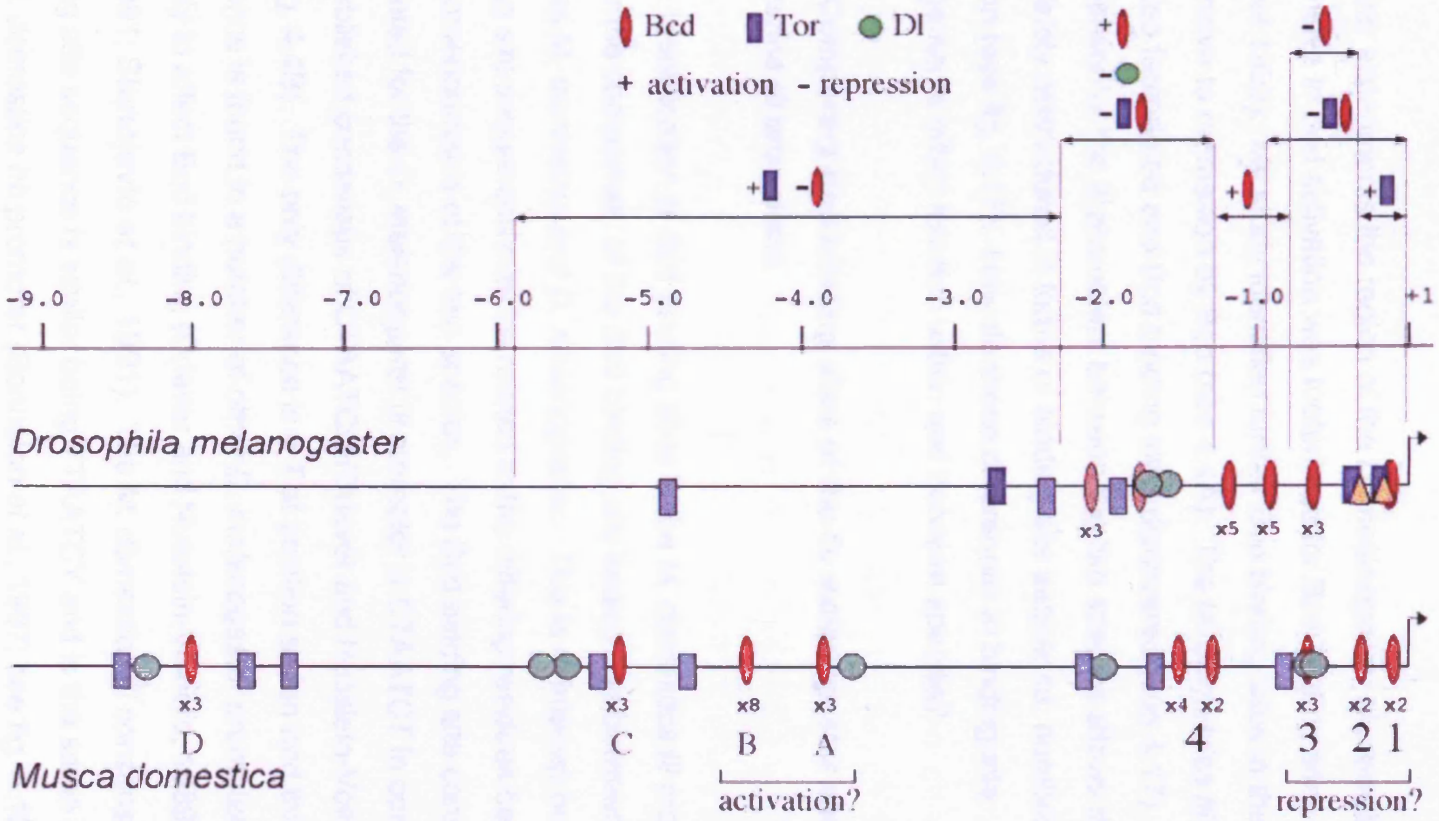


Figure 4.18 A comparison of Bcd binding sites identified in the *D. melanogaster* and *M. domestica tll* promoters including the *D. melanogaster tll* regulatory regions previously identified (Liaw and Lengyel, 1992). Other transcription factor sites have been identified in the *D. melanogaster tll* promoter and these are shown. The sites are as follows: red ovals: Bcd, blue rectangles: Tor-RE, green circles: Dl and yellow triangles: TTK69. Striped sites are putative binding sites. The numbering below sites refers to the number of binding sites present at that location. The regulatory regions identified by Liaw and Lengyel are represented above the scale with arrows marking the position of each r egion. The symbols above represent the transcription factors involved in the regulation of that region.

4.3 Discussion

4.3.1 Footprinting the *tll* promoters of *D. melanogaster* and *M. domestica*

In earlier experiments the region of the *D. melanogaster tll* promoter that is responsive to Bcd activation was footprinted for Bcd binding sites (Liaw and Lengyel 1992). My study identified further Bcd binding sites in the region responsive to repression by Bcd (see 4.4A). The *M. domestica tll* promoter was also footprinted and Bcd binding sites discovered (see 4.17). Comparison of the *tll* promoters between the two species shows that they are completely restructured in terms of binding site sequence, number and position (see fig. 4.17). How do these differences in binding site arrangements affect function within and between species?

4.3.2 Comparing Bcd binding sites of the *D. melanogaster* and *M. domestica tll* promoters

Identification of Bcd binding sites in the *M. domestica tll* promoter can further the comparison of the Bcd binding site sequences between the species *M. domestica* and *D. melanogaster.* This is of interest because the binding site sequences may be related to the differing residues between the Bcd homeodomains of the two species. The Bcd binding site consensus calculated for the *D. melanogaster tll* promoter is CTAATCT in comparison to the published consensus of CTAATCC (Driever and Nusslein-Volhard, 1989; see fig. 4.4B). The only difference is a T at position seven and this alternate sequence is found in a number of other *D. melanogaster* promoters so is unlikely to affect Bcd binding (Driever and Nusslein-Volhard, 1989; Hoch *et al.,* 1991; Stanojevic *et al.,* 1991). The *M. domestica tll* consensus Bcd binding site sequence is similar being TTAATCY and is the same as that of the *M. domestica hb* promoter (Bonneton *et al.*, 1997; see fig. 4.15). Thus, there is a consistent difference in the preferred base at position one between *D. melanogaster* and *M. domestica* (see table 4.4).

	D. melanogaster	*M. domestica*
hb	YTAATCC	TTAATCY
tll	CTAATCT	TTAATCN

Table 4.4 The consensus Bcd binding site sequences of *D. melanogaster* and *M. domestica*

It has been suggested that the variation in the Bcd consensus binding sites between *D. melanogaster* and *M. domestica* could be related to the five differences in the Bcd homeodomain between the species (Bonneton *et al.*, 1997; see 9.2.2). However, there is more variation in the Bcd binding site sequences within species than between species. Common deviations from the consensus were observed in both species, such as an A at position one or two and G at position five. Bcd binds by the formation of direct contacts between residues in the homeodomain and specific bases in the binding site (see 1.9; Ades and Sauer, 1995; Tucker-Kellogg *et al.*, 1997). Binding of non-consensus sites is possible if the bases present can make new contacts with the homeodomain to replace those lost. For example, it has been shown that arg54, which makes a contact with the A at position three in a TAAT core binding site can make a new contact with the G of a TAAG core binding site by a rotation of the arg54 side chain (Dave *et al.*, 2000). Some bases are never seen at a particular position in the binding site presumably because of their failure to make favourable contacts with the Bcd homeodomain; for example, C at position three or five and G at position two and four.

The affinity of Bcd for non-consensus sites can be increased due to the ability of Bcd protein to bind DNA in a cooperative manner (see 1.9). This means that a site with a good match to the consensus found near a site with a poor match to the consensus will greatly enhance the affinity of Bcd for the latter site such as the X2 and X3 sites in *D. melanogaster hb* respectively (Yuan *et al.*, 1999). Interestingly, many Bcd sites with a poor match to the consensus are found close to another site with a good match to the consensus in both *tll* promoters (see fig. 4.3 and 4.14). Therefore a site that is considered weak because of a poor match to the consensus binding sequence may actually be more important to promoter function than a site with a good match to the consensus binding site sequence. This suggests

that the arrangement of Bcd binding sites is more significant to promoter function than the sequence of individual binding sites.

4.3.3 Differences between Bcd sites found in activating and repressing regions

Functional studies have shown that Bcd is involved in both the activation and repression of the *D. melanogaster tll* promoter (Liaw and Lengyel, 1992). The mechanism by which Bcd represses *tll* expression is not fully understood but is dependent on Tor (see 1.12). Bcd function may be better understood by comparing the binding sites involved in activation and repression. Therefore the region responsive to Bcd mediated repression of *tll* was footprinted and a further four sites identified (see fig. 4.4A). It should be noted that additional sequences resembling Bcd binding sites were present in the repressing region but these were not protected.

The density of sites in the repressing region is much lower than those in the activating region (see 4.17). Although three of the sites, in the repressing region, were close enough to allow for interactions between Bcd proteins only one pair was arranged favourably for cooperative binding (see 4.2.1; Yuan *et al.,* 1999). In addition, the repressing sites had poor matches to the Bcd consensus sequence (see 4.2.1 and table 4.1). Therefore, Bcd protein would not have a great affinity for these sites and this would impact on any cooperativity of binding between the sites. In contrast, the activating region binding sites with a poor match to the consensus are positioned closely to other sites and this allows co-operative interactions to occur which enhances binding to these sites (see 4.3.5; Burz *et al.,* 1998). Similarly, the *hairy* stripe 7 and *knirps* 64 elements contain binding sites with poor matches to the Bcd consensus binding site yet both elements are responsive to Bcd. In both cases the Bcd binding sites are arranged suitably to allow cooperative interactions between Bcd proteins (Riddihough and Ish-Horowicz. 1991; Rivera-Pomar *et al.*, 1995).

Repression of *tll* expression at the anterior cap is dependent on high levels of Bcd protein as increasing or decreasing Bcd dosage shifts the repression to the posterior or to the anterior respectively (Pignoni *et al.*, 1992). It is possible that the low density, poor consensus sites, may only be bound at

the higher levels of Bcd protein. One proposed form of repression called "quenching" is mediated by the binding of repressor proteins close to activating sites (<100 bp) to prevent activation from these sites (Gray and Levine, 1996). Since the Bcd repressing region corresponds to the Tor derepressing region, it is possible Bcd is quenching the Tor regulated derepression in the anterior cap (see 4.3.7).

A further possibility is that, binding site D1 (see fig. 4.4A), which is located 97 bp from the transcription start, could be directly repressing activation. A Bcd protein bound at this position could be interfering with the factors that bind to the basal promoter perhaps in combination with Tor phosphorylated dSAP18 (see 1.12; Gray and Levine, 1996; Zhu *et al.*, 2001). Although these possibilities are intriguing, whether the difference in density observed between sites in the repressing and activating regions are functional or coincidental will not be known until further functional analysis is carried out.

4.3.4 Comparing the arrangement of Bcd binding sites in the *D. melanogaster* and *M. domestica tll* promoters

The ability of Bcd to bind cooperatively means that the spacing between sites can directly influence the function of Bcd (Ma *et al.*, 1996; Burz *et al.*, 1998; Yuan *et al.*, 1999). Therefore, comparison of the organization of the sites in *D. melanogaster* and *M. domestica* is important for the understanding of differences in Bcd regulation of the two *tll* promoters. The Bcd binding sites of the *D. melanogaster tll* promoter cover a region of 1.6 kb and are known to function in activation and repression (Liaw and Lengyel, 1992; see fig. 4.17). In contrast the identified *M. domestica* Bcd binding sites are spread over 8.5 kb of sequence.

The arrangement of Bcd binding sites in *M. domestica* may give an indication of their function, such as the spacing between sites as seen in *D. melanogaster* (see 4.3.3). Some of the most favourable spacing for cooperative binding is seen in region B and this arrangement should result in a strong affinity for Bcd protein (Yuan *et al.*, 1999; see fig. 4.16). Region A is located close to B and sites in this region also exhibit spacing favourable for

cooperative binding. These regions combined can be predicted to have a high activating potential for the *tll* promoter.

The sites located more proximally to the transcription start are less dense, although they are always within 100 bp of a neighbouring Bcd site. As with *D. melanogaster* additional potential Bcd binding site sequences were observed in this region but they were not protected in the footprinting experiment. This indicates the difference in spacing of footprinted sites is real and that this region could function in repression as discussed for *D. melanogaster* binding sites above. Indeed results from transgenic analysis presented later suggest the repression of *tll* at the anterior tip is located in this region (see 7.3.2).

In conclusion, the *M. domestica* Bcd binding sites are organised locally in clusters much like *D. melanogaster*. In *M. domestica* these clusters are spread throughout a much larger regulatory region. The expansion of regulatory sequences has also been observed for the *M. domestica hb* promoter. This promoter is also completely restructured with respect to the *D. melanogaster hb* promoter but is functionally equivalent to it (Bonneton *et al.*, 1997). Therefore, the expansion of the *tll* regulatory sequences in *M. domestica* is not unusual and is unlikely to have affected the function.

4.3.5 Further factors affecting the regulation of *M. domestica tll*

There are other factors that may influence the regulation of *tll* by Bcd, such as the proposed synergistic relationship between Bcd and Hb (Simpson-Brose *et al.*, 1994; see chapter 1). Whether Hb enhances *tll* activation by Bcd could be investigated by determining *tll* expression observed in an *hb* mutant. Indeed, there are many putative Hb binding sites within the *tll* promoter.

To establish if the other binding sites identified in the *M. domestica* sequence are functional a similar strategy to the one described in the work for Bcd would be necessary. Putative Tor-RE sites were observed, some of which were found in the Bcd dependent repressing region (see 4.3.3), as well as potential Dl binding sites and the relevance of these will be discussed later (see 7.3.2). It is likely that some of the GAGA binding sites observed in the promoter sequence are functional as it is a ubiquitously expressed

transcription factor, which has been shown to be involved in activation and repression at a number of promoters (Liaw *et al.*, 1995). Interestingly GAGA factor interacts with dSAP18, a Bcd co-factor and as some GAGA sites overlap the Tor-RE, GAGA may be involved in the interaction between Bcd and Tor (Espinas *et al.*, 2000).

4.4 Summary

The Bcd binding sites identified here are suggestive of the presence of both activating and repressing regions responsive to Bcd within the *tll* promoter of *M. domestica*. This is indicative of conservation of function of the promoter sequences between *D. melanogaster* and *M. domestica*. These results are in accordance with the demonstration that *tll* and *bcd* expression and *bcd* function are conserved between the species (this work; Sommer and Tautz, 1991; Shaw *et al.*, 2001). All of which indicates conservation of Bcd regulation of *tll* in this species. However, it is apparent that the *M. domestica tll* promoter has become completely restructured in terms of general organization of Bcd binding sites from the *D. melanogaster* promoter. How do two homologous regulatory regions become completely unalignable whilst maintaining function? It is the aim of the next chapter to examine this question with a study of the microevolution of the promoter sequences.

Chapter 5 Intra-specific analysis of the *tll* gene in *M. domestica*

5.1 Introduction

5.1.1 Comparing the regulatory sequences of *D. melanogaster* and *M. domestica tll*

The *M. domestica tll* gene has a conserved pattern of expression with respect to *D. melanogaster tll* (see 3.3.4). Identification of Bcd binding sites *in vitro* (see 4.3.1) and an *in vivo* transgenic analysis (see 7.2.2) suggest that sequences upstream of the *tll* coding region in *M. domestica* are responsible for *tll* regulation. However, the *M. domestica tll* regulatory sequences are unalignable with those of *D. melanogaster tll* (Liaw and Lengyel, 1992). How have the regulatory sequences retained the same function but diverged at the sequence level? This question is central to understanding how *cis*-regulatory sequences evolve.

5.1.2 Analysis of non-coding sequence evolution

To discover how mutational mechanisms and selection shape the evolution of DNA sequences a number of statistical tests have been developed (for review see Kreitman, 2000). However, these tests have focused on coding sequences because the genetic code gives an inherent structure to the coding sequence and constrains sequence change because of the requirements for translation fidelity and protein function. Importantly these features of coding regions have enabled the development of a theory of the evolution of these sequences (Bergman and Kreitman, 2001). Conversely, non-coding regions are generally unalignable except between closely related species and there is a lack of knowledge about the rules governing the structure and subsequent evolution of functional non-coding sequences (Stern, 2000).

The few comparative studies of the evolution of *cis*-regulatory sequences have necessarily been between closely related species and the alignments suggest a pattern of conserved blocks of functional sequence interspersed

amongst divergent sequences (Kim, 2001; Bergman and Kreitman, 2001). For example, a study of the evolution of the conserved *eve* stripe 2 promoter between closely related *Drosophila* species indicated that whilst the binding site sequences could still be aligned the sequences between sites varied in length and sequence composition (Ludwig *et. al.*, 1998). Even between these closely related *Drosophila* species the sequences have diverged to the point where individual mutations cannot be distinguished. Similarly, a comparison of the *hb* *cis*-regulatory regions between *D. melanogaster* and *D. virilis* identified conserved blocks of sequences containing Bcd binding sites yet the remaining sequence had no similarity (Lukowitz *et al.*, 1994). These studies suggest that the non-functional parts of the *cis*-regulatory sequences may be under little selection against change, whilst the functional sequences such as binding sites are conserved.

5.1.3 An intra-specific analysis of the *M. domestica hb* gene

An intra-specific analysis should demonstrate the first steps in the divergence of two species by revealing the variation present within a species. This variation is only the first step in the evolution of a sequence and should therefore provide a snap-shot of the mutational processes which lead to the evolution of new species and the rate of these events in non-coding DNA sequences.

The *M. domestica hb* P2 promoter is functionally equivalent to the *D. melanogaster hb* P2 promoter, but the sequences are unalignable. An intra-specific analysis of the *M. domestica hb* P2 promoter was carried out with six *M. domestica* strains (McGregor *et.al.*, 2001). The sequences of the six strains were aligned and the results are shown in table 5.1.

	Length (bp)	Indel	Base substitution
P2 Promoter	764	12	24
5'UTR	321	5	8
Coding	1550	4	73

Table 5.1 Differences between strains in the *M. domestica hb* gene

Knowledge of the evolution of coding sequences can help the determination of the evolution of non-coding sequences. For example, both the promoter region and the 5' UTR of the *hb* gene have a lower rate of substitution than the rate of change at third position bases in the coding region and this indicates that there is selection against sequence changes in both the promoter and 5' UTR regions to maintain function (Kimura, 1983). Indeed, none of the Bcd binding site sequences contain base substitutions. However it is possible that slippage and gene conversion events, which can remove polymorphism occur more frequently in the non-coding regions because of the different selection constraints.

As expected there are fewer insertion/deletion (indel) events in the coding sequences than the non-coding sequences since the majority of length changes would result in a frame shift of the coding region. None of the indels in the promoter are found within Bcd binding site sequences and presumably have little or no effect on the function of these sequences (Mcgregor *et al.*, 2001). These results suggest that both base substitutions and indels can be tolerated in *cis*-regulatory sequences but are not observed within binding sites.

5.1.4 Identifying mechanisms that generate sequence variation

Promoter sequence comparisons show that whilst blocks of functional sequence remain conserved the surrounding sequences diverge considerably, but what is responsible for this sequence variation? As can be seen from the results of the *M. domestica* intra-specific variation analysis, sequences are subject to mutation in the form of point mutations and also insertion and deletion events. The latter are a result of a number of mechanisms, such as unequal crossing-over, gene conversion and slippage. DNA slippage events are thought to be about 100 times more frequent than point mutations (Schlotterer and Tautz, 1992; Hancock, 1995; Schug *et. al.*, 1998; Harr *et. al.*, 2000). Therefore, it is likely that slippage events have played a significant role in the evolution of non-coding regions of the genome (Hancock, 1995).

The slippage of DNA sequences during replication is caused by repetitive DNA sequences (Levinson and Gutman, 1987; Schlotterer and Tautz, 1992; Hancock, 1996). These sequences consist of short repeated motifs of two to four bases in length and are also known as simple sequences (Tautz *et. al.*, 1986). There are also cryptically simple sequences that are di-, tri- or tetranucleotide motifs interspersed one with another, for example:

simple sequence: TAATAATAATAATAA
cryptically simple sequence: TAATGTATAAGTAATAA

5.1.5 Identifying simple sequences

Regions of simple sequence may be more prone to DNA slippage processes (see 5.1.4; Hancock, 1995; Shug *et al.*, 1998). To identify such regions the SIMPLE34 programme was devised and can detect both simple and cryptically simple sequences (Hancock and Armstrong, 1994; based on the SIMPLE programme, Tautz *et. al.*, 1986). The programme also determines the frequency of simple motifs present in a sequence, hence enabling the comparison of the frequency and type of simple motifs between sequences (Hancock and Armstrong, 1994; see 5.2.2).

Various studies suggest that regions of high simplicity are subject to frequent turnover events (Tautz *et. al.*, 1986; Schlotterer and Tautz, 1992; Schug *et. al.*, 1998) but can evidence for this be seen in sequences such as the rapidly evolving promoter sequences of *M. domestica* and *D. melanogaster*? Analysis of the *M. domestica hb* intra-specific comparison revealed that the indel events observed in the coding and 5' UTR sequences were significantly located in regions of high simplicity. In the promoter region indels were found equally in regions of high and low simplicity (McGregor *et. al.*, 2001). These results are suggestive of a link between simple sequences and indel events.

5.1.6 Aims

The evolution of gene sequences is dependent on both the mutation and recombination rate, which can vary throughout the genome (Tautz and Negro, 1998). Therefore any studies of non-coding sequence mutation and divergence should contain information from a number of sequences within the genome. The evolution of a *cis*-regulatory sequence will also depend on the functional constraints of the sequence. Therefore the aim of this chapter was to compare the evolution of the *M. domestica tll* gene with the *hb* gene. This involved an intra-specific comparison of *tll* gene sequences and a simple sequence analysis.

5.2 Materials and methods

5.2.1 Sequencing of the *tll* gene from *M. domestica* strains

The five *M. domestica* strains sequenced were: Cardiff, Millan, Rentokil, Scott, White and Zurich (see 2.1.2). The regions sequenced were the proximal promoter region, the 5'UTR, the intron and the coding region (see fig. 5.1). The same primer sets were used with every strain and to minimize sequencing errors each region was sequenced from two independent PCRs (see 2.2.3).
The sequences were aligned with the 'working' *tll* sequence (Rutgers strain) using ClustalW. The numbers of indels and base polymorphisms were counted for each region.

5.2.2 Simple analysis of the *tll* gene sequences

The SIMPLE34 programme assigns a simple score to each base in a sequence by analysing the repeats present in a 64 bp sliding window (see fig. 5.2). For a particular base 'C' positioned at the centre of the sliding window which includes the 32 bp 5' and 3' the score is calculated as follows: taking 'C' as the first base of either a tri- or tetranucleotide repeat, the 64 bp window is scanned for repeats

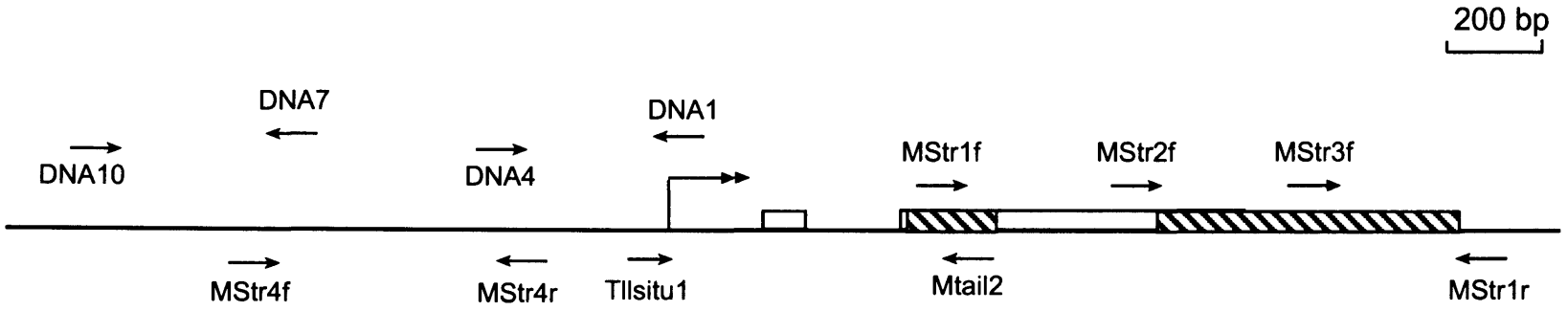


Figure 5.1 Primers used for sequencing of the 5 strains of *M. domestica tll*.
Primer position and orientation are shown in relation to the *tll* transcript (see 5.2.1). The double arrowhead indicates the transcription start site and the boxes the coding region (striped areas indicate functional domains).

of the tri- and tetranucleotide motifs. Each time the trinucleotide motif is repeated the base is awarded a score of one and for each tetranucleotide repeat a score of three (Hancock and Armstrong, 1994; see fig. 5.2).

This procedure gives a simplicity profile of the sequence, averaging the scores of all nucleotides in the sequence produces a simplicity factor (SF). To obtain a relative measure of simplicity within the sequence [relative simplicity factor - RSF] a simplicity factor is calculated for ten random sequences of the same length and base composition as the test sequence. The simplicity factor is then divided by the mean of the ten 'random' simplicity factors to give the RSF. If the RSF is greater than one, then the test sequence contains a higher simplicity than expected by chance. Many sequences have been shown to have an RSF higher than one, including both coding and non-coding regions (Tautz *et. al.*, 1986; Hancock, 1995; 1996; Hancock *et. al.*, 1999; McGregor *et. al.*, 2001).

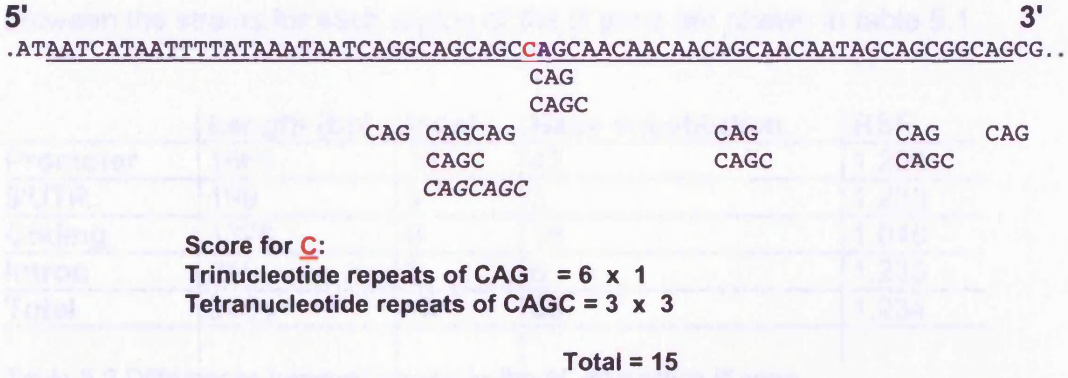


Figure 5.2 Calculating the simplicity score for the central C (red) is shown
The 64 bp window includes the bases underlined and the repeats of the tri and tetranucleotide commencing at the central C are shown below. The score for this base (15) and how it is calculated is shown in bold. Once the score has been calculated the window slides 3' by one base and the score for the A (blue) is calculated. The programme uses a window of 64 bp and searches for only tri- and tetranucleotide motifs to reduce background noise whilst optimizing signal match (Tautz *et. al.*, 1986). The programme excludes overlapping motifs such as the tetranucleotide shown in italics, only allowing the motif to score once.

The SIMPLE34 programme identifies the motifs present in the sequence, which are responsible for the high level of simplicity at a given point. To identify

the motifs that are present at a proportion significantly higher than expected, a significance value (S) for each motif is calculated using the equation:

$S = 1-(fe/fo)$

Where fo is the observed frequency of that motif in the test sequence and fe is the mean frequency in the ten random sequences. A sequence is considered to be present at a significantly high level when $S = \geq 0.9$, i.e. when fo is ten times greater than fe.

5.3 Results

5.3.1 Comparison of the *tll* sequences between the strains of *M. domestica*

The *tll* gene was sequenced in six strains of *M. domestica* and the sequences aligned (see Appendix 2). The numbers of indels and base substitutions seen between the strains for each region of the *tll* gene are shown in table 5.1.

	Length (bp)	Indel	Base substitution	RSF
Promoter	1669	15	43	1.204
5'UTR	199	3	2	1.233
Coding	1326	0	15	1.049
Intron	210	1	6	1.233
Total	3404	19	66	1.234

Table 5.2 Differences between strains in the *M. domestica tll* gene

The rate of base substitution in the *tll* coding region is lower than that seen in the non-coding regions. In the coding region all base polymorphisms were at silent sites. Nine out of the fifteen substitutions in the *tll* coding region were found outside the functional domains that make up 70 percent of the protein (see fig. 3.4) and this is significant ($\chi^2 = 6.14$, $p<0.05$ at 1 d.f.). No indels are observed within the coding region, which suggests there are constraints on length changes within the protein and selection against mutations altering the reading frame.

In the promoter sequence none of the identified Bcd binding sites are interrupted by indels or base substitutions. One polymorphism is found close to

Bcd binding site 10 and results in a change from ATAATCTC to ATAATCTT, so it is unlikely to abolish binding at this site (see fig. 4.9). The absence of polymorphisms in the binding sites is statistically non-significant but this is due to the large region over which the binding sites are spread.

5.3.2 Simple sequence analysis of the *M. domestica tll* gene

A dotplot analysis of the *M. domestica tll* gene revealed there was a number of small repeats present throughout the sequence and the coding and non-coding regions were shown to share some of the short motifs present (see fig. 5.3). The SIMPLE34 programme furthers this comparison by identifying similarities and differences that cannot be detected by a simple alignment. In doing this, the programme generates relative values for the simplicity of a sequence, which allows comparison between sequences.

The *M. domestica tll* sequence from the Rutgers strain was split into promoter, 5'UTR, coding and intron regions and the individual parts analysed (see Appendix 3 for SIMPLE34 data). The results of these analyses are presented in figures 5.4, 5.5 and 5.6. The RSF values calculated for each region are shown in table 5.1. The coding region has the lowest RSF of 1.049, which is not significantly different from one (see fig. 5.4). The sequences with the highest simplicity within the coding region are found between the DNA binding domain and the ligand-binding domain. Only one motif, ATTG, is present at a significant level in the coding region.

Of the non-coding sequences, only the promoter region was significantly simple ($p<0.01$). Despite both the intron and 5'UTR having the highest RSF values the sequences were too short for the RSF to be significant. The binding sites in the promoter region were not found within simple sequences (see fig. 5.5). Two motifs were present at a high frequency within the sequence, TTTT and TAAA, the former identified at 14 different positions. Therefore this TTTT motif was found to be present at significantly high levels, S = 1.

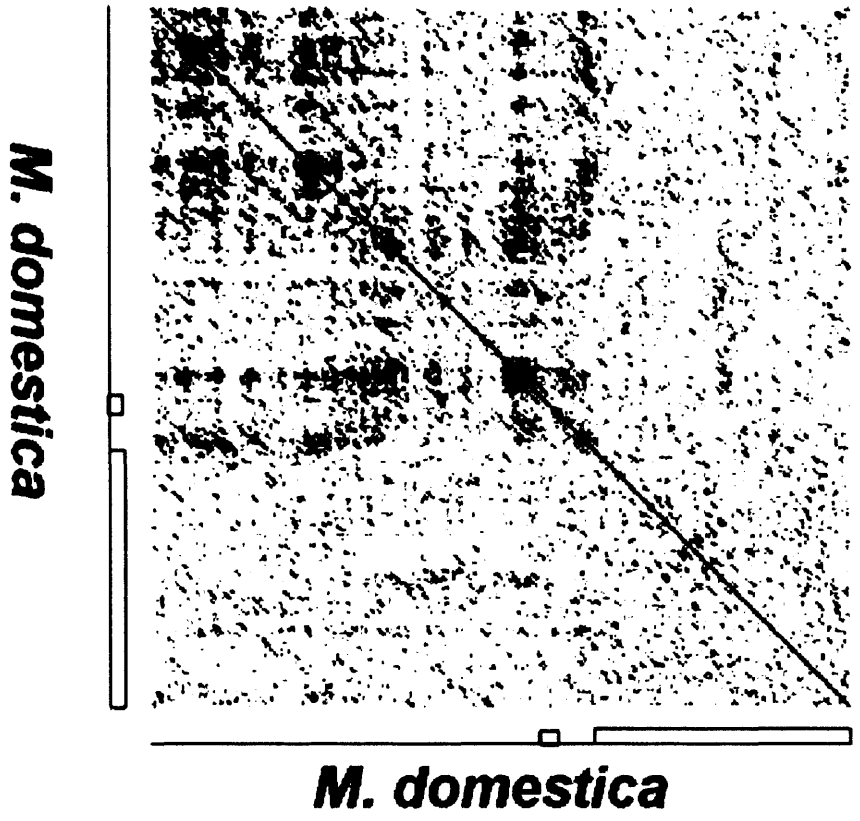


Figure 5.3 Dot-plot of an intra-specific sequence comparison of *M. domestica tll*. This was generated using the dot plot programme (see 2.2.12), a window of 35 bp and a stringency of 18 bases matching. The sequence position is indicated by the transcript structure shown, the black boxes indicating the coding region.

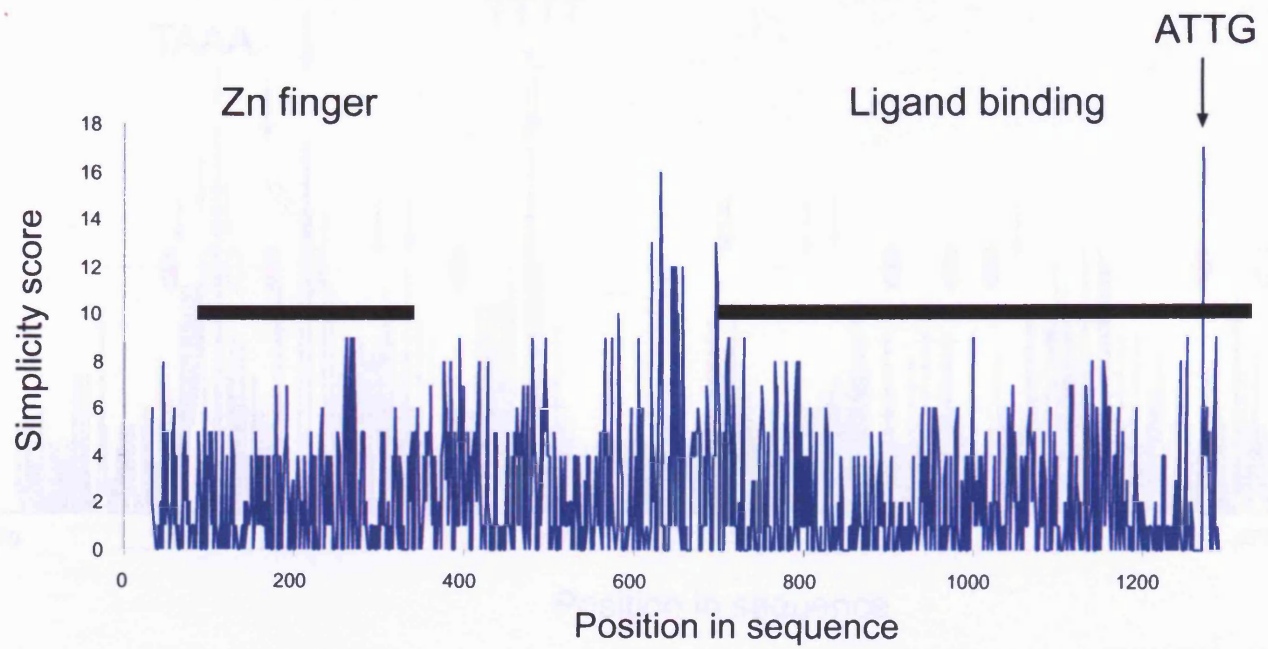


Figure 5.4 A graph showing the simplicity scores of the *tll* coding region. The sequence position relative to the translation start site is shown on the y-axis. The simplicity score is shown on the x-axis. The structure of the coding region is indicated by the black line, with the boxes representing the ligand-binding and zinc finger domains. The position of the one 'significant' repeat (see text) is indicated by the arrow above the plot.

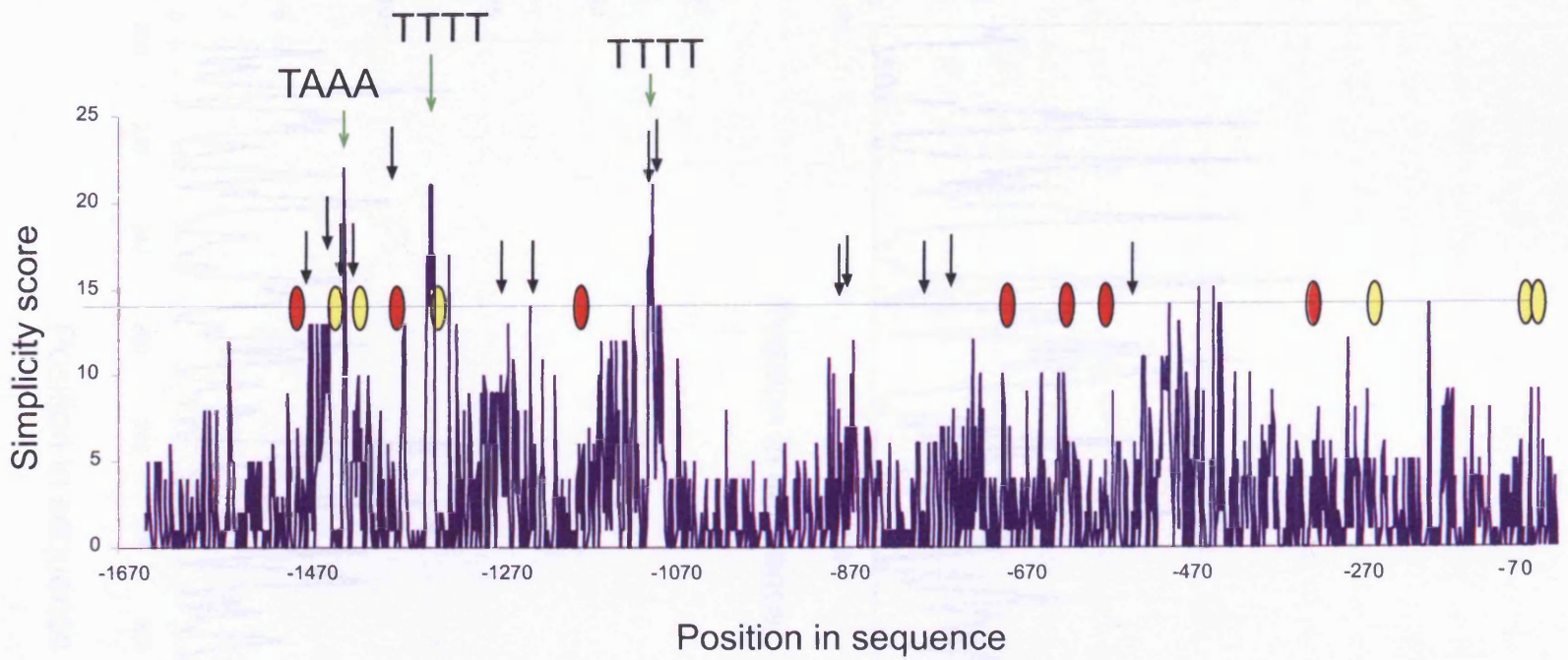


Figure 5.5 A graph showing the simplicity scores of the *tll* promoter region. The sequence position relative to the transcription start site is shown on the y-axis. The simplicity score is shown on the x-axis (note the scale is different to fig. 5.4). The position of the binding sites are indicated by ovals. Red sites indicate core binding site sequences and yellow sites represent non-core binding site sequences. The positions of the 'significant' repeats (see text) are indicated (green arrows). The black arrows show the positions of the indels between the strains of *M. domestica*.

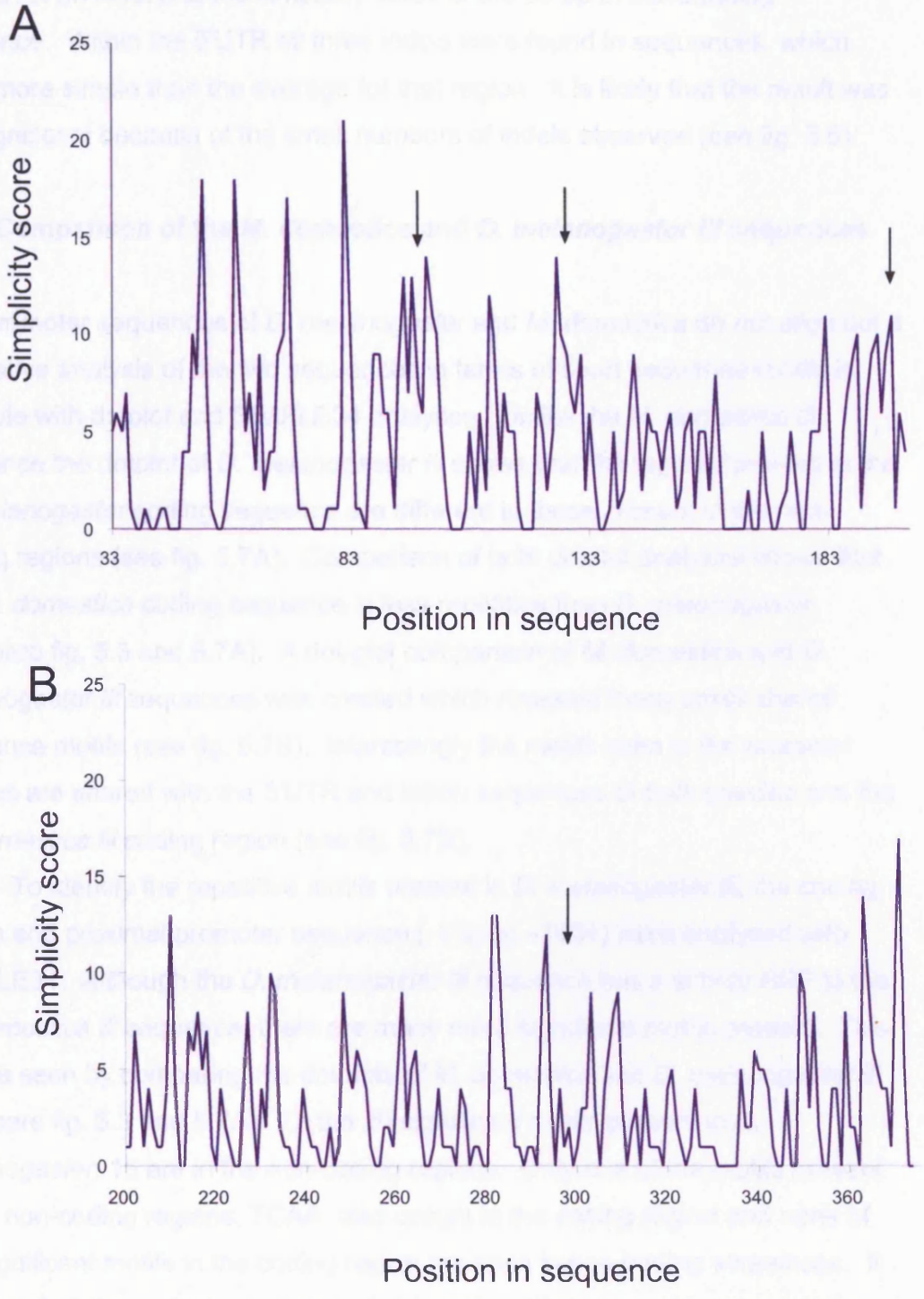


Figure 5.6 A graph showing the simplicity scores of the *tll* 5'UTR (A) and intron (B). The sequence position relative to the transcription start site is shown on the y-axis. The simplicity score is shown on the x-axis (note the scale is different to fig. 5.4). The black arrows show the positions of the indels between the strains of *M. domestica*.

None of the regions of the *tll* gene showed a significant link between the location of an indel and the simplicity value of the 50 bp of surrounding sequence. Within the 5'UTR all three indels were found in sequences, which were more simple than the average for that region. It is likely that the result was not significant because of the small numbers of indels observed (see fig. 5.6).

5.3.3 Comparison of the *M. domestica* and *D. melanogaster tll* sequences

The promoter sequences of *D. melanogaster* and *M. domestica* do not align but a qualitative analysis of the two sequences in terms of short sequence motifs is possible with dotplot and SIMPLE34 analyses. Unlike the *M. domestica tll* sequence the dotplot of *D. melanogaster tll* shows that the repeats present in the *D. melanogaster* coding sequence are different to those present in the non-coding regions (see fig. 5.7A). Comparison of both dotplot analyses shows that the *M. domestica* coding sequence is less repetitive than *D. melanogaster* (compare fig. 5.3 and 5.7A). A dot-plot comparison of *M. domestica* and *D. melanogaster tll* sequences was created which revealed many small shared sequence motifs (see fig. 5.7B). Interestingly the motifs seen in the promoter regions are shared with the 5'UTR and intron sequences of both species and the *M. domestica tll* coding region (see fig. 5.7B).

To identify the repetitive motifs present in *D. melanogaster tll*, the coding region and proximal promoter sequence (-1 bp to –1634) were analysed with SIMPLE34. Although the *D. melanogaster tll* sequence has a similar RSF to the *M. domestica tll* sequence, there are many more significant motifs present. This can be seen by comparing the dotplots of *M. domestica* and *D. melanogaster tll* (compare fig. 5.3 and 5.7A). Of the 20 significant motifs present in *D. melanogaster*, 15 are in the non-coding regions. Only one of the motifs present in the non-coding regions, TCAA, also occurs in the coding region and none of the significant motifs in the coding region are seen in non-coding sequences. It appears that the mutually exclusive division of motifs between the two regions

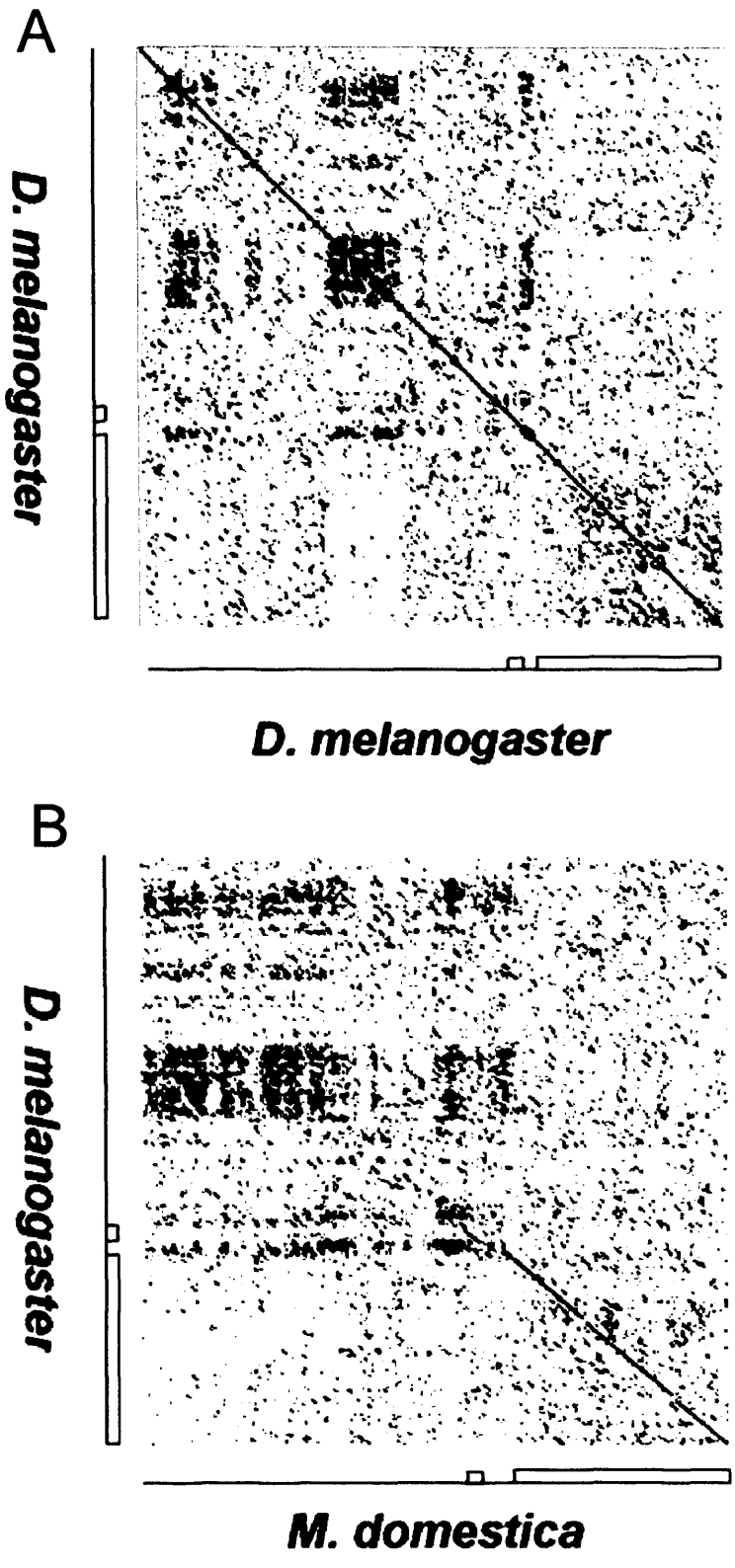


Figure 5.7 Dot-plots of an intra-specific sequence comparison of *D. melanogaster tll* (A) and an inter-specific sequence comparison between *D. melanogaster* and *M. domestica tll* (B). These were generated using the dot plot programme (see 2.2.12), a window of 35 bp and a stringency of 18 bases matching. The sequence position is indicated by the transcript structure shown, the black boxes indicating the coding region.

reflects the AT content of the repeats as those found in the non-coding regions have a higher AT content (see Appendix 3).

The small number of significantly occurring repeats in the *M. domestica tll* sequence limits the comparison between the species. However the two motifs seen in the *M. domestica* promoter are also found in the *D. melanogaster* promoter sequence.

5.4 Discussion

5.4.1 Evolutionary analysis of *M. domestica tll* gene

To begin to understand the mechanisms by which a non-coding functional region evolves a study of the *M. domestica tll* gene sequence was carried out. An intra-specific comparison of six *M. domestica* strains led to an estimate of the amount and type of polymorphism present in the different regions of the *tll* gene. Analysis of the sequence composition using the SIMPLE34 programme, identified regions that were highly repetitive for short motifs. Unlike the *hb* gene analysis there was no significant relationship between indels and the simple sequences.

5.4.2 Rates of base substitution in the *tll* gene in *M. domestica*

The relative level of substitution in the different regions of the *tll* gene reflects the amount of constraint experienced by functional versus non-functional sequences. The greater number of base substitutions in the non-coding regions could be due to the lower proportion of functionally relevant sequences. Base substitution in the 5' UTR is much lower than that seen at the coding sequence third position, which indicates there is selective constraint on this sequence, or that base substitutions have been obscured by other mutational processes such as slippage. It is possible that the 5'UTR is experiencing constraints because of its role in the transcription and translation of the *tll* gene. The intron sequence has a rate of polymorphism similar to that at the coding region third position, which suggests that there is little constraint on the sequence of the intron. This correlates with the small number of functionally significant intronic sequences required, such as the splice sites and the branch point. The substitution rate for the promoter region is similar to the intron rate. Although this sequence is functional the high substitution rate can be explained by the structure of the cis-regulatory sequences. These consist of short functionally constrained binding

site sequences interspersed within much longer regions of DNA with presumably no sequence constraint. No polymorphism is observed within the Bcd binding sites, this agrees with the expectation that the sequences are under selective constraint because of their function in the regulation of *tll*.

The level of polymorphism observed at the *tll* gene locus is less than that seen at the *hb* locus, but the difference is not constant between the coding and non-coding regions of each gene (McGregor *et. al.*, 2001, see table 5.2). In particular, the *hb* coding region has a much greater rate of substitution than the *tll* coding region, on average every seventh silent site is polymorphic in *hb* compared to every thirtieth in *tll*. In addition, six of the polymorphisms in the *hb* coding region are non-synonymous, although none of these occur in the functional domains (McGregor *et. al.*, 2001).

	Length (bp)	**Indel**	**Base polymorphism**	**RSF**
Total for tll	3404	19	66	1.204 P 1.049 C
Total for hb	2678	23	111	1.628 P 2.152 C

Table 5.3 Comparison of polymorphisms in the *tll* and *hb* genes. P refers to promoter and C to coding region.

The difference in the level of polymorphism between these two genes could be influenced by the recombination rate in each region. It is known that the recombination rate correlates positively with polymorphism and that recombination rates vary across the genome in *D. melanogaster* and *H. sapiens* (Tautz and Nigro, 1998; Kreitman, 2000; Jensen *et. al.*, 2002; Aquadro *et. al.*, 2001). The mechanisms that cause this phenomenon are likely to be general and therefore should be the same for *M. domestica* (Kreitman, 2000).

The levels of polymorphism seen in the *M. domestica tll* gene are more similar to *M. domestica hb* than to those observed for *D. melanogaster hb* (Tautz and Nigro, 1998). A study of 12 strains of *D. melanogaster* found eighteen polymorphic sites over a 3.3 kb sequence of the *hb* gene. Comparison with the genome averages for *D. melanogaster* showed that this level of polymorphism at

the *hb* locus was low. However, *hb* is found in a region of low recombination so the low rate seen for *D. melanogaster hb* was proposed to be a result of a selective sweep in this region of the genome (Moriyama and Powell, 1996; Tautz and Nigro, 1998).

5.4.3 Sequence length changes in the evolution of the *tll* gene promoter in *M. domestica*

As transcription factors work together to regulate gene expression, the distance between binding sites can be functionally significant. It has been proposed that whilst mutations in the sequences between binding sites are neutral, there is selection to maintain spacing between sites involved in cooperative interactions (Ludwig *et. al.*, 1998). However, there was no evidence for compensatory length changes, such as both an insertion and deletion between two binding sites in the *tll* promoter of any strain in the intra-specific comparison. Therefore, the ability to observe step by step the compensatory length changes in cis may not be possible from sequence comparison alone.

A less informative but feasible approach is to look for constraints on length changes between sites of more distantly related species. This has been attempted for the *even-skipped* stripe 2 enhancer in *Drosophila*. It was shown that despite extensive changes in the sequences lying between the binding sites, the distances between sites were conserved between *Drosophila* species (Ludwig *et. al.*, 1998). Although compensatory change could not be demonstrated at the sequence level, *in vivo* transgenic assays showed that stabilising selection was acting to maintain function through compensatory changes within the promoters (see fig. 1.3; Ludwig *et. al.*, 2000). However, the *M. domestica* and *D. melanogaster tll* promoters are too diverged to attempt this kind of comparison as equivalent binding sites cannot be identified (see chapter 9 for further discussion).

Although compensatory changes could not be identified, it is interesting that in both the *M. domestica tll* and *hb* promoters the indels seen between

closely positioned binding sites are all small, one or two bases in length. This could indicate the need to keep these sites close together for cooperative interaction between Bcd molecules (Mao *et al.*, 1994). However, this observation may be an artifact of the small number of large indels observed and the probability of one of these falling in the short region between closely spaced sites.

5.4.4 The relationship between sequence length changes and simplicity

The relative contribution of different mutational mechanisms such as unequal crossing-over and slippage, which result in indel events, is unknown. However, as some of the indels observed in the *tll* gene are found at the end of mononucleotide runs it is likely slippage mechanisms are involved in the formation of these. Slippage occurs at repetitive sequences and so the simplicity of a sequence can influence the rate of slippage events (Hancock, 1995). The indels in the *M. domestica* sequence were not significantly correlated with highly repetitive sequences. However, a significant correlation was seen between the indel position in the *hb* coding region and sequences with high simplicity ($p<0.01$; Mcgregor *et. al.,* 2001). In comparison, no indels were seen in the *tll* coding region and the difference between the simplicity of the two coding regions is striking with an RSF of 2.152 for *hb* in comparison to 1.049 for *tll* (see table 5.2). Indeed, the frequency of indels in the *M. domestica hb* gene is twice that of *tll* and this correlates with the greater simplicity of the *hb* sequence.

A comparison of 17 genes between *T. castaneum* and *D. melanogaster* has indicated a link between sequence repetition, divergence and length. The *T. castaneum* genes are an average of 30% shorter than *D. melanogaster* due to the virtual absence of trinucleotide repeats in *T. castaneum* genes along with a lower degree of internal repetition (simplicity; Schmid and Tautz, 1999). The difference in the RSF values and number of repeats for each gene comparison varied from one gene to the next. This suggested that the differences in RSF values and repetition were not due to a general difference between genomes but

reflected the differing functional constraints of the genes (Schmid and Tautz, 1999).

The indels found in the *hb* coding region were in glutamine and histidine repeats. The length of these repeats in the *hb* gene varies between different species (McGregor *et. al.*, 2001). Tll however does not have any such amino acid repeats and is a much shorter gene. Therefore, the relative simplicity of the sequence of *hb* and *tll* may contribute to their differences in length and conservation. Hb is only 66% conserved between *M. domestica* and *D. melanogaster* in comparison to Tll at 83%.

The fewer indels seen in the *tll* gene in comparison to *hb* could be related to the local recombination rates as well as the sequence simplicity. In addition, the evidence from the comparison of *D. melanogaster* and *T. castaneum* genes suggests functional constraints can affect the simplicity of a sequence. This could affect the difference in simplicity of the *tll* promoter compared to *hb* (see table 5.3). The *tll* promoter is bound by multiple regulatory factors as opposed to the *hb* P2 promoter that contains only Bcd binding sites (Liaw and Lengyel, 1992; Driever and Nusslein-Volhard, 1989). These factors do not behave independently and it is likely therefore that there are more constraints against length changes between these different sites in the *tll* promoter.

5.4.5 Sequence content, simplicity and evolution

The calculation of simplicity allows for a comparison of sequences between *M. domestica* and *D. melanogaster* even though they cannot be aligned. Dot-plot analysis shows that the two sequences do share many small repetitive motifs between the coding regions and also between the non-coding regions. Only three motifs were present at significant levels in *M. domestica tll* yet the dot-plot of *M. domestica* indicates the presence of many repeats. A possible explanation is that the motifs are scrambled so do not occur in the same density as in simple sequences. This scrambling has been previously observed for *T. castaneum hb* sequences (Hancock *et. al.*, 1999).

The motifs identified in *D. melanogaster tll* reveal a bias towards AT rich motifs in the non-coding regions and AT poor motifs in the coding regions. Excepting one motif these are mutually exclusive. This suggests that the two regions have different selective constraints on sequence composition (Akashi, 2001). The *M. domestica* motifs are all AT rich and a distinction between motif composition throughout the gene is not observed. The analysis reveals that the promoter regions of the two species share the same AT rich motifs. Interestingly, of the motifs identified in the *D. melanogaster tll* promoter and *M. domestica hb* P2 promoter, those proximal to the coding region contain C and G, but those more distal contain only A and T (McGregor *et. al.*, 2001). AT rich motifs are known to be much more prone to slippage than motifs containing G or C and most of the indels identified in the *M. domestica tll* gene intra-specific comparison involved runs of T or A (Schlötterer and Tautz, 1992; Hancock, 1995). Indeed, 10 out of 15 indels in the *M. domestica tll* promoter fall in the distal 700 bp of sequence where the AT content is over 70%. Interestingly, all known *M. domestica* non-coding sequences are AT rich and the genome is estimated at three and a half to five times larger than *D. melanogaster* (Crain *et al.*, 1976). All the sequence expansion identified in *M. domestica* has been in non-coding regions, as seen for *tll*, *hb*, *bcd* and other genes (Bonneton *et. al.*, 1997; Shaw *et. al.*, 2001; J Clayton personal communication). Perhaps slippage mechanisms have played a substantial role in this genome expansion. Indeed, evidence for the involvement of slippage in genome expansion has been found in a diverse group of species (Hancock, 1996).

5.4.6 Summary

This study of intra-specific polymorphism shows that the rate of divergence of the *M. domestica tll* gene is different to that of *M. domestica hb* and provides a more extensive analysis of the evolution of non-coding sequences. The *tll* gene is less polymorphic than *hb*, which could be related to functional constraints of the *tll*

gene. In particular, in the promoter region the constraints may be higher than those seen in *hb* due to the greater regulatory complexity of the *tll* promoter.

There is some evidence that the simplicity of a sequence is involved in the turnover of that sequence. However, a link between these simple sequences and indel events identified in the intra-specific analysis was unproven.
The results suggest a model of non-coding evolution by which the non-functional sequences are evolving rapidly whilst the functional sequences are conserved. The question of the function of the two *tll* promoters and their interaction with the Bcd protein will be presented in the next two chapters.

Chapter 6 Functional analysis of the Bcd-*tll* promoter interaction between *M. domestica* and *D. melanogaster*

6.1 Introduction

6.1.1 The evolution of the Bcd-*tll* promoter interaction between *D. melanogaster* and *M. domestica*

The regulation of *tll* by Bcd has been conserved between *D. melanogaster* and *M. domestica* over the 100 MY since they last shared a common ancestor (Beverley and Wilson, 1984). However, the interaction has diverged extensively at the molecular level. The Bcd homeodomain has five differences between the species and the *tll* promoter sequences are unalignable (Bonneton *et al.*, 1997; this work). The lack of sequence conservation of regulatory regions has been observed between species in comparisons of promoters with conserved function (for review see Tautz, 2000). The generation of sequence variation is a result of mutation and genomic turnover events (Schug *et. al.*, 1998). How these variants are tolerated and spread through a population whilst the interaction is maintained is not known (Ohta and Dover, 1983; Dover and Flavell, 1984).

One possible mechanism is molecular co-evolution, which has been described (see 1.5). Briefly, a *cis*-regulatory sequence variant is produced as a result of mutation and genomic turnover events. This variant may then increase in the population until there is compensatory selection for a transacting factor, which is better adapted to interact with the variant sequence (Ohta and Dover, 1983; Dover and Flavell, 1984). Thus whilst the function has been maintained the molecular basis of the interaction has changed.

A result of co-evolution is that the components of an interaction will diverge between two species. Eventually the components of an interaction will be incompatible with those of another species. It is possible that the changes observed in the Bcd-*tll* promoter interaction are due to the co-evolution of the components of the interaction within *M. domestica* and *D. melanogaster*.

6.1.2 Bcd protein function, the role of binding affinity and cooperativity

Bcd is a morphogen, activating different genes at different concentrations. Activation of Bcd targets depends on the number of binding sites for the Bcd protein present in the *cis*-regulatory module, the sequence of the binding sites and the arrangement of sites (Berleth *et al.*, 1988; Rivera-Pomar *et al.*, 1995; Burz *et al.*, 1998). These features are important because they affect the affinity of Bcd for the *cis*-regulatory module.

The affinity of Bcd for a binding site is determined by the concentration of Bcd at which the binding site is occupied. The sequence of a binding site directly affects the affinity of Bcd for that site. In general, sites with a greater match to the consensus binding site have a high affinity for Bcd and are known as strong sites, for example the *D. melanogaster* consensus TCTAATCC (Ma *et. al.*, 1996; Burz *et. al.*, 1998). Conversely, sites with a poor match to the consensus have a low affinity and are known as weak sites.

Analysis of Bcd regulation of target promoters has shown that binding site sequences are not the only factor affecting activation. The arrangement of sites is also important because Bcd can bind DNA cooperatively (see 1.9). The cooperative interaction between Bcd molecules involves direct contact between the surfaces of each Bcd protein. Binding sites have to be within 100 bp of each other and arranged in such a way that the correct contacts between the two proteins can be made for cooperative binding to occur (Ma *et al.*, 1996). The cooperativity of an interaction can be measured by the increase in protein concentration over which a module changes from an unbound state to having all sites occupied. The smaller the change in concentration of protein necessary for binding all sites the greater the cooperativity of the interaction. Cooperative interactions between sites affect the binding affinity of sites. This means that Bcd may bind with a high affinity to a site with a poor match to the consensus if the site is arranged cooperatively with another site. For example, the *D. melanogaster hb* promoter X2 site has a weak consensus sequence, but when

deleted results in the greatest loss of gene activation out of all sites in the *hb* promoter (Yuan *et. al.*, 1999; Shaw *et. al.*, 2002).

6.1.3 Testing the binding ability of the Bcd proteins

Measuring the affinity and cooperativity of an interaction between Bcd and a target promoter is a good indication of the strength of that interaction (Ackers *et al.*, 1983). To measure the affinity and cooperativity of an interaction *in vitro* a band shift assay can be used (Mao *et al.*, 1994). Bound promoter fragments can be distinguished because they travel more slowly through a gel (see fig. 6.1 and 6.2). At increasing concentrations of Bcd protein the numbers of binding sites occupied will increase and the fragments will move more slowly. A graph can then be plotted of the concentration of Bcd against the occupancy of binding sites (see fig. 6.3). The affinity of the interaction is taken as the concentration of Bcd where half of the sites are occupied and is called the affinity constant (Km). If an interaction is cooperative the line produced is sigmoidal, thus the steeper the gradient of the line the more cooperative the interaction. The slope of the line is used to calculate the cooperativity, which is known as the Hill coefficient (n).

A band shift assay can be used to calculate affinity and cooperativity of an interaction using the components of two different species. This provides a test for the co-evolution of an interaction (Bonneton *et. al.*, 1997; Shaw *et. al.*, 2002).

6.1.4 *D. melanogaster* and *M. domestica* Bcd affinities for single binding sites

In vitro experiments with Bcd protein and single binding sites of the *hb* promoters of *D. melanogaster* and *M. domestica* showed that *Drosophila* Bcd bound to all sites with a greater affinity than did *Musca* Bcd (Shaw, 1998, Shaw *et al.*, 2002). However, the difference in Bcd binding affinity between the two species varied from site to site, from a nearly five times difference for the *M. domestica hb* G1

site to an almost equivalent affinity for the *D. melanogaster hb* A1 site (Shaw 1998, Shaw *et al.*, 2002).

Analysis of the individual binding sites showed that there was no correlation between binding site sequence and binding affinity with either protein, indicating that the sequences flanking the binding site were affecting binding (Shaw 1998, Shaw *et al.*, 2002). For instance, the local DNA topology may influence binding by Bcd (Breiling *et al.*, 2001).

The results of these earlier assays indicate that it is unwise to characterise the overall strengths of a promoter in terms of its individual binding sites. A more realistic indication of an interaction between a transcription factor and promoter would be a measurement of the affinity for a group of binding sites in their species-specific configurations.

6.1.5 Aims

The differences present within the Bcd proteins of *D. melanogaster* and *M. domestica* may have altered the binding properties of the two proteins (see fig. 1.7). The accompanying changes observed between the composition of the *tll* promoters may be the result of co-evolutionary changes in the interaction. An *in vitro* band shift assay can compare the binding ability of the two Bcd proteins and reveal differences in the compatibility of the components of the two interactions.

6.2 Materials and Methods

6.2.1 Band-shift assays using the tll promoters

For the band-shift assay a region was chosen from each of the *D. melanogaster* and *M. domestica tll* promoters. The *D. melanogaster tll* promoter fragment contained five Bcd binding sites from the region responsive to activation by Bcd (bases –1194 (Dmtll5) to –1026 (Dmtll6), Pignoni *et al.*, 1990; Bcd binding sites 4-8 of Liaw and Lengyel 1992 or D10-14 this study, see fig. 4.3). The *M.*

domestica tll promoter fragment (bases –1743 (DNA12) to –1356 (DNA9) of *M. domestica tll*, this work, see fig. 4.5) included Bcd binding sites 10 to 13 (see fig. 4.10).

The promoter fragments were radioactively labeled and mixed with dilutions of *D. melanogaster* and *M. domestica* Bcd homeodomain-GST fusion proteins and then the complexes separated by gel resolution (see 2.2.10).

The fraction of DNA bound (in all complexes) was determined by comparing band intensities of bound and free DNA complexes. The average results of 4 separate reactions were fitted to the equation: Ybar = Kn*Xn / (1 + Kn*Xn) using the program DATAFIT (see 2.2.10). Ybar refers to the fraction of the DNA, which is bound by protein. K is the apparent equilibrium constant, X is the concentration of Bcd protein and n is the Hill coefficient (see 2.2.10.2).

6.3 Results

6.3.1 Testing the Bcd-*tll* promoter interactions of *D. melanogaster* and *M. domestica*

Band-shift assays were carried out using *tll* promoter fragments from either species in conjunction with the *D. melanogaster* Bcd and *M. domestica* Bcd homeodomains (see fig. 6.1 and 6.2). At increasing concentrations of Bcd homeodomain several protein-DNA complexes were observed indicating stoichiometric binding to multiple sites in the *D. melanogaster* and *M. domestica tll* promoters (see fig. 6.1 and 6.2, C1-4). The binding affinity (Km) and cooperativity (n) of each protein for both promoters were then calculated (see table 6.1, fig. 6.3 and 2.2.10). The complete results of the band shift assays are shown in Appendix 4.

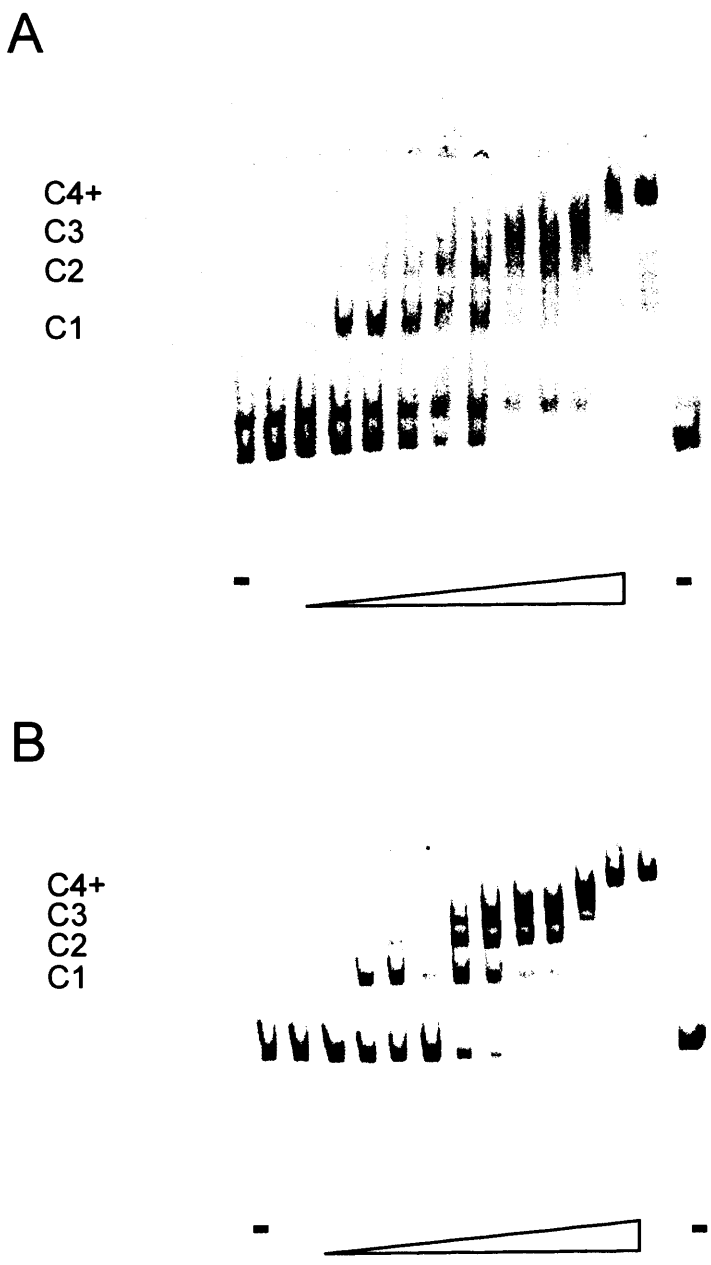


Figure 6.1 Band shift assays of *D. melanogaster* Bcd binding *tll* promoter fragments
A. *D. melanogaster* Bcd bound to the *D. melanogaster tll* promoter.
B. *D. melanogaster* Bcd bound to the *M. domestica tll* promoter.
The first and last lanes are control reactions to which no Bcd protein was added. In these two lanes the DNA fragments are diffusing in an unbound state.
Lanes 2 to 13: Bcd protein ranging from 10pm to 100nm active protein was added to the reactions (the triangle indicates the increasing amounts of Bcd protein). With increasing Bcd concentration complexes of increasing molecular weight are seen, these refer to the increased occupancy of the binding sites within the DNA fragment (C1 to C4+).

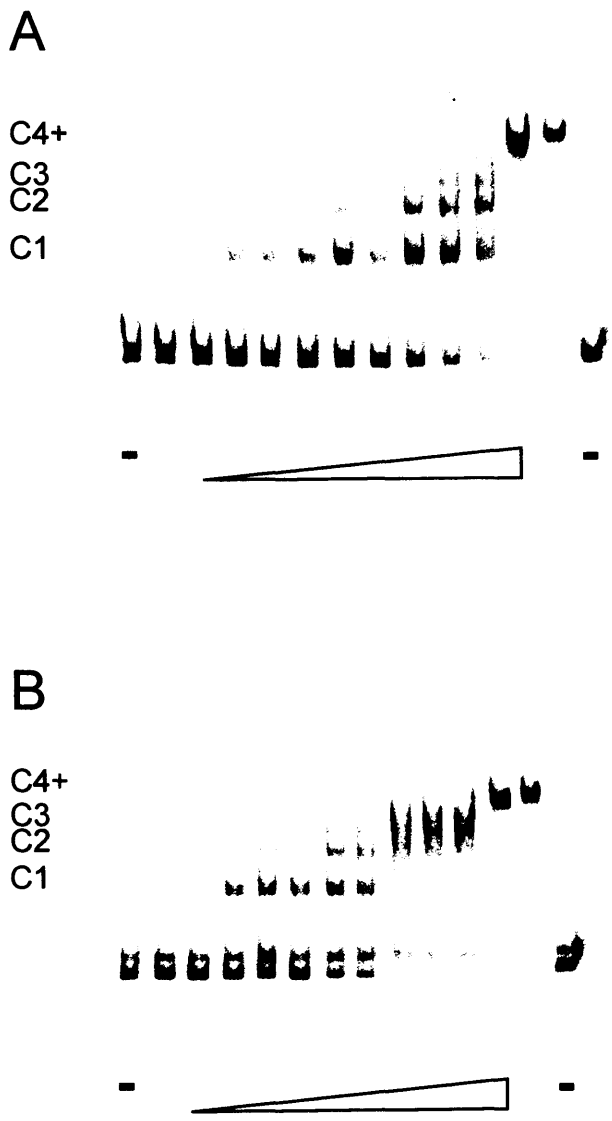


Figure 6.2 Band shift assays of *M. domestica* Bcd binding *tll* promoter fragments A. *M. domestica* Bcd bound to the *M. domestica tll* promoter B. *M. domestica* Bcd bound to the *D. melanogaster tll* promoter The first and last lanes are control reactions to which no Bcd protein was added. In these two lanes the DNA fragments are diffusing in an unbound state. Lanes 2 to 13: Bcd protein ranging from 10pm to 100nm active protein was added to the reactions (the triangle indicates the increasing amounts of Bcd protein). With increasing Bcd concentration complexes of increasing molecular weight are seen, these refer to the increased occupancy of the binding sites within the DNA fragment (C1 to C4+).

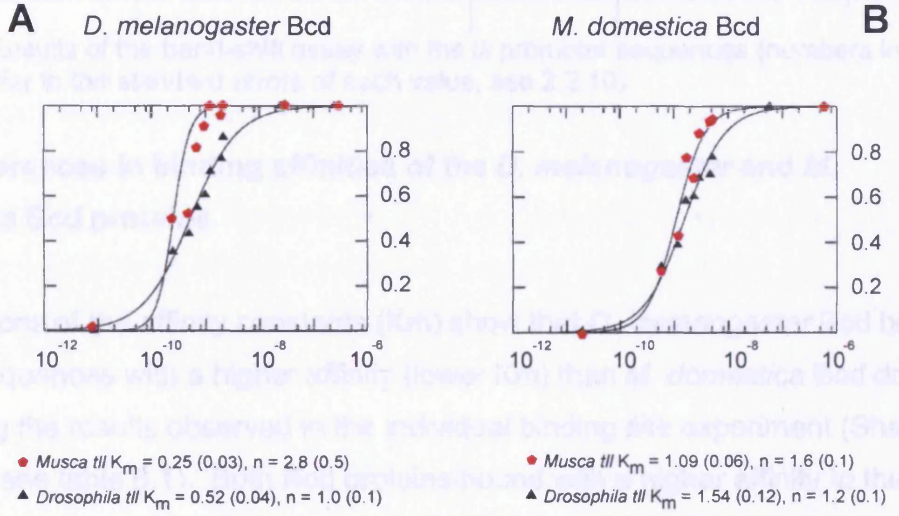


Figure 6.3 Graph showing the binding affinity curves of *D. melanogaster* and *M. domestica* Bcd with the *D. melanogaster* and *M. domestica tll* promoters. *Drosophila* Bcd (**A**) and *Musca* Bcd (**B**) *tll* promoter DNA binding curves were generated by curve fitting to mean data points. Fractional saturation (y-axis) is plotted against active molar Bcd concentration (x-axis). The fitted values of affinity constant (K_m) and Hill coefficient of cooperativity (n), with the standard errors of each value in parentheses are shown underneath the curves.

	***D. melanogaster* Bcd**		***M. domestica* Bcd**	
	Km	n	Km	n
***D. melanogaster tll* promoter**	0.52 (0.04)	1.0 (0.1)	1.54 (0.12)	1.2 (0.1)
***M. domestica tll* promoter**	0.25 (0.03)	2.8 (0.5)	1.09 (0.06)	1.6 (0.1)

Table 6.1 Results of the band-shift assay with the *tll* promoter sequences (numbers in brackets refer to the standard errors of each value, see 2.2.10)

6.3.2 Differences in binding affinities of the *D. melanogaster* and *M. domestica* Bcd proteins

Comparisons of the affinity constants (Km) show that *D. melanogaster* Bcd binds to both sequences with a higher affinity (lower Km) than *M. domestica* Bcd does, supporting the results observed in the individual binding site experiment (Shaw et al., 2002; see table 6.1). Both Bcd proteins bound with a higher affinity to the M*tll* promoter (see table 6.1; compare red circles (*M. domestica tll*) to black triangles (*D. melanogaster*) in fig. 6.3A and B).

There was variation in the amount of cooperative binding that occurred in the interactions (see table 6.1; n values greater than one indicate cooperativity). There was no cooperative binding to the *D. melanogaster tll* promoter by *D. melanogaster* Bcd and very little by *M. domestica* Bcd. In contrast both proteins bound cooperatively to the *M. domestica* promoter (compare red circles to black triangles in fig. 6.3A and B). Interestingly, *D. melanogaster* bound with a much greater cooperativity than *M. domestica* Bcd to the *M. domestica tll* promoter (see table 6.1).

The low cooperative binding values for the *D. melanogaster* promoter suggests that the *D. melanogaster* promoter is less optimally arranged for cooperative binding than the *M. domestica* one (see table 6.1). The greater cooperativity of the *M. domestica* promoters contributed to the lower affinity constants of both proteins binding the *M. domestica tll* promoter (see fig. 6.3). This could account for the very low Km of *D. melanogaster* Bcd for the *M. domestica tll* promoter.

6.4 Discussion

6.4.1 Comparing the Bcd-*tll* promoter interaction between *D. melanogaster* and *M. domestica*

A band shift assay was used to test the binding affinity and cooperativity of *D. melanogaster* and *M. domestica* Bcd proteins for the *tll* promoters of both species. The results show that *D. melanogaster* Bcd has a greater binding affinity for DNA than *M. domestica* Bcd. The cooperativity observed for both proteins was greater on the *M domestica tll* promoter which may be a consequence of selection for binding sites arranged to enhance cooperativity between Bcd proteins in *M. domestica*.

6.4.2 What could be causing the difference in binding affinity of the two proteins?

The Bcd homeodomain recognises and binds to DNA sequences. Therefore, the homeodomain residues, which vary between the two species are candidates for the difference in binding affinity of the two Bcd proteins (Bonneton *et al.*, 1997; see 6.1.2). Although none of these residues contact the DNA directly, the general, high conservation in homeodomain sequences indicates the importance of the structure of this domain for binding. Investigations into the binding ability of Bcd have shown that the flexibility for binding different sequences is due to variable rotation of amino acid side chains to make specific contacts with the DNA. These results show how important the structure of the homeodomain is in terms of shape and charge so that rotations of the side chains are possible to bind non-consensus sites (Dave *et al.*, 2000, Zhao *et al.*, 2000). Therefore, it is interesting that the changes in the homeodomain between the two species are not conservative, altering the charge of each of the residues concerned. For example, at position 11 of the homeodomain in *D. melanogaster* there is a polar residue (ser) and in *M. domestica* there is a non-polar residue (ala), whereas at

positions 28, 29 and 30 in *D. melanogaster* there are non-polar residues but in *M. domestica* polar residues (see fig. 1.7). Changes such as these could alter the shape of the protein and so may be a result of selection.

In both species the sequences flanking the binding sites affect Bcd binding affinity (Shaw *et. al.*, 2002). Therefore, it is possible that sequences outside the homeodomains, particularly those that maintain the shape of the Bcd protein, are also important for binding. Support for this idea comes from a binding experiment with a chimaeric Ftz protein in which the Ftz homeodomain was replaced with the Bcd homeodomain (Zhao *et al.*, 2000). This protein was able to recognise consensus Bcd binding sequences but not other known Bcd binding sites. Other evidence from methylation protection experiments shows that when Bcd binds certain sites, sequences outside the consensus site are being contacted (Dave *et al.*, 2000). Therefore, it is likely the structure, shape and charge of the whole Bcd protein is important in its ability to bind DNA.

The binding affinity of each protein is affected by the ability to bind cooperatively, this is certainly the case with *D. melanogaster* Bcd binding to the *M. domestica tll* promoter (see table 6.1; red circles in fig. 6.3A). The homeodomain and flanking sequences are necessary for cooperative binding so the changes observed in these regions between the Bcd proteins could also be affecting cooperativity (Yuan *et al.*, 1996; Burz and Hanes, 2001). Indeed, one of the residues in the homeodomain that differs between *D. melanogaster* and *M. domestica* is involved in cooperative binding (Burz and Hanes, 2001).

6.4.3 Comparing the Bcd-*hb* promoter interaction between *D. melanogaster* and *M. domestica*

A similar band shift assay was used to compare the binding affinity for *D. melanogaster* and *M. domestica* Bcd proteins for the *hb* promoters of both species (Shaw *et. al.*, 2002). The results of this experiment showed that *D. melanogaster* Bcd had a higher binding affinity than *M. domestica* Bcd on both *hb* promoters (see table 6.2). This agreed with the *tll* promoter and other band

shift experiments (Shaw *et. al.*, 2002; this work). Both Bcd proteins had a higher affinity for the promoter of the same species (see table 6.2). The cooperative values showed the opposite trend with the Bcd proteins having a higher cooperativity with the promoter of the other species. *D. melanogaster* Bcd bound to the *M. domestica hb* promoter with the highest cooperativity as was the case with the *tll* promoters.

In comparison to the results with the *tll* promoters, the affinity constants were lower with the *hb* promoters. As the *hb* promoters are activated in a region of lower Bcd protein concentration it is conceivable that the *hb* promoters will have evolved to have a higher affinity for Bcd than the *tll* promoters. Indeed, although the Bcd proteins bound to the *D. melanogaster tll* promoter with almost no cooperativity this was not the case for the *D. melanogaster hb* promoter.

	***D. melanogaster* Bcd**		***M. domestica* Bcd**	
	Km	n	Km	n
***D. melanogaster hb* promoter**	0.20 (0.01)	1.7 (0.1)	0.53 (0.04)	1.9 (0.3)
***M. domestica hb* promoter**	0.22 (0.02)	2.2 (0.1)	0.34 (0.02)	1.4 (0.1)

Table 6.2 Results of the band-shift assay with the *hb* promoter sequences (numbers in brackets refer to the standard errors of each value, see 2.2.10.2)

6.4.4 The Bcd proteins of each species bind the promoters in different ways

D. melanogaster Bcd has a weak binding ability in comparison to many other DNA binding proteins and the *M. domestica* Bcd protein has been shown to be even weaker (Shaw *et al.*, 2002; Ma *et. al.*, 1999). Therefore, it is possible that in *M. domestica* the promoters could have been selected to allow for a greater amount of cooperative binding between Bcd proteins, due to the weak binding ability of the *M. domestica* Bcd protein. Of course the reciprocal argument could be true, that because the *M. domestica* promoter sequences were more suited to cooperative binding the selection for the *M. domestica* Bcd protein to maintain a

strong binding ability was relaxed or selected against (Small *et al.*, 1991 and 1996).

The selection for cooperative arrangements of Bcd binding sites in *M. domestica* promoters could be a response to the gradient of Bcd protein in these embryos. The *M. domestica* embryos are twice the size of *D. melanogaster* and the Bcd protein has to diffuse over a larger distance. As a result the *M. domestica* Bcd protein could be at a lower concentration than *D. melanogaster* Bcd in equivalent regions of the egg. A more cooperatively arranged promoter would compensate for the lower concentration of Bcd protein.

The fact that *D. melanogaster* Bcd has a strong cooperative binding ability, which is seen when the protein is assayed with *M. domestica tll* or *hb* promoters, supports the possibility of the ancestral Bcd having a strong cooperative binding ability (Shaw *et al.*, 2002; see table 6.1). However, this ability is masked in the *D. melanogaster* Bcd- *D. melanogaster* promoter interactions tested, since the *D. melanogaster* promoters do not appear to be optimally arranged to enhance cooperative binding. A cooperative interaction has the most noticeable effect when protein concentrations are the limiting factor (Burz *et al.* 1998). Therefore, it may be easier to demonstrate the latent cooperative binding ability present in the *D. melanogaster* Bcd protein on a promoter such as the *kni* 64 box promoter which is activated at low concentrations of Bcd (Rivera-Pomar *et al.*, 1995; Burz *et al.*, 1998). These promoters which are activated at low concentrations of Bcd are possibly maintaining selection for the cooperative binding ability of *D. melanogaster* Bcd.

6.4.5 Comparison of Bcd affinity for within-species and between-species promoter interactions

The *in vitro* assay can reveal differences between the within-species and between-species Bcd-promoter interactions of *D. melanogaster* and *M. domestica*. It would be expected that any sequence differences seen in the protein or promoter regions between species would not be detrimental to the

interaction within a species (see 6.1.1; Dover and Flavell, 1984). Meanwhile, the changes observed could disrupt the between species interaction because the components aren't co-evolving. Over time as differences between the species accumulated the interaction should become weaker and this would be reflected as a decreased binding affinity. The band shift assay would then reveal these differences in binding affinity. The results for *M. domestica* Bcd agreed with the expectation of a higher binding affinity for the *tll* promoter of the same species (see fig. 6.3B). However, the *D. melanogaster* Bcd protein had a higher affinity for *M. domestica tll* (see fig. 6.3A). The *D. melanogaster* Bcd also gave approximately the same value for *M. domestica hb* as the *D. melanogaster hb* promoter (see table 6.1; Shaw *et. al.*, 2002). How can these results be reconciled with the expectation of a higher affinity within-species interaction?

One explanation is that in each species the Bcd proteins are maintaining the same level of activation of target genes by slightly different means. For example the *M. domestica* Bcd protein has a lower affinity for DNA but the promoter sequences appear to be arranged to enhance cooperative binding. Whereas, *D. melanogaster* Bcd has a greater affinity for DNA and a less cooperative interaction with *D. melanogaster* promoters. Thus, when the components from each species are mixed, the *D. melanogaster* Bcd-*M. domestica tll* promoter interaction has a greater affinity than expected because of a combined greater affinity and co-operativity. In agreement with this the *M. domestica* Bcd-*D. melanogaster tll* interaction has the lowest affinity. The interactions involving the *hb* promoters show a similar pattern but to a lesser degree.

Evidence from a yeast transcriptional assay supports the idea that the Bcd-*hb* promoter interactions have evolved differently between the species and that Bcd proteins are more compatible with the promoter of the same species (Shaw *et al.*, 2002). Such assays demonstrated that activation of the *D. melanogaster hb* promoter was greater with *D. melanogaster* Bcd than with *M. domestica* Bcd. Activation from the *M. domestica hb* promoter was equal with both Bcd proteins. This indicates that *M. domestica* Bcd is more compatible with

the *hb* promoter of the same species (Shaw *et. al.*, 2002). Activation by *D. melanogaster* Bcd from the *M. domestica hb* promoter was greater than expected but could be explained by the greater binding affinity of *D. melanogaster* Bcd and the more cooperatively arranged *M. domestica hb* promoter. However, these results could have been affected by the difference in the preferences of the yeast transcription complex from those of insects (Hanes *et al.*, 1994; McGregor, 2002).

6.4.6 Limitations of the *in vitro* analysis

In analyzing the results of the *in vitro* assay it has been assumed that a lower Km is an indication of selection for maintenance of an interaction (Shaw *et al.*, 2002). This assumption may be a suitable test for an interaction that functions as a simple on/off switch, such as in a signal transduction cascade. However, Bcd functions as a morphogen, activating a number of genes at different concentrations of Bcd protein throughout the egg (Driever and Nusslein-Volhard, 1988b). The binding affinity of Bcd combined with the ability to bind cooperatively is therefore required to interact at many different levels. This means that the requirement for Bcd binding will be different for the *tll*, *hb* and other promoters that are activated further to the posterior (Pignoni *et al.*, 1992; Struhl *et al.* 1989; Rivera-Pomar *et al.* 1995). Over-activation of Bcd targets can be just as detrimental as insufficient levels of Bcd (Gibson, 1996; Janody *et. al.*, 2001). Indeed over expression of Bcd in the anterior can be lethal to embryos because of ectopic expression of Bcd activated genes in the posterior (Namba *et. al.*, 1997). Therefore, the use of parameter such as strength of binding may be misleading as a measure of the co-evolution of an interaction within different species.

Another consideration is that a strong binding affinity of Bcd for either *hb* or *tll* promoters may not be necessary as these genes are activated in the anterior of the egg where there are high levels of Bcd protein (Driever and Nusslein-Volhard, 1988b). Therefore, a promoter which responds to low

concentrations of Bcd such as the *kni* promoter may function as a better indicator of the co-evolution of an interaction between the two species (Burz *et al.*, 1998).

6.4.7 Redundancy of Bcd functions in early development

The interaction between Bcd and promoter sequences is further complicated by the redundancy and buffering of interactions in early development (see 1.3; Carroll *et al.*, 2001; Waddington, 1942). It has been shown that an up to 30% difference in the amount of Bcd can activate the same correct expression domain of *hb*, although the precise mechanism for this remains unknown (Houchmandzadeh *et al.*, 2002). The regulation of most early developmental genes involves multiple transcription factors, which work in concert to drive the correct expression patterns of the gene (Arnone and Davidson, 1997). For instance activation of *tll* involves both Bcd and Tor (Pignoni *et al.*, 1992; see 7.3.3). Although only Bcd is necessary for activation in *tor*$^-$ mutants, the stripe is less well defined and it has been shown that *torso* activity increases the activation by Bcd (Pignoni *et al.*, 1992). There is also evidence that *hb* activation is increased by autoactivation by Hb protein and that Hb can replace much of the Bcd functions in the anterior (Simpson-Brose et al., 1994; Wimmer *et. al.*, 2000). Therefore small changes in the affinity or cooperativity of Bcd, which appears to be the case between these closely related species may not affect the correct expression of its target genes. Indeed, both *M. domestica bcd* and *hb* transgenes can rescue function in a *D. melanogaster* maternal *bcd* and zygotic *hb* mutant, respectively, albeit at a lower level than a *D. melanogaster* transgene (Bonneton *et. al.*, 1997; Shaw *et. al.*, 2002). Thus, the changes observed between species may be a result of neutral drift and eventual compensatory changes in *cis* or due to selection in a different type of Bcd interaction. These different evolutionary trajectories for the divergence of the Bcd interaction will be discussed further (see chapter 9).

6.4.8 Summary

The *in vitro* assay shows that *D. melanogaster* Bcd has a greater affinity for DNA than *M. domestica* Bcd and that the different combinations of protein and DNA can result in quite different Bcd binding affinities. This illustrates how very flexible the interactions between a transcription factor and promoter can be especially when factors such as cooperative binding ability and multiple binding sites are introduced. These results hint at the features of the Bcd protein that enable it to activate target genes along the anterior-posterior axis.
Although these studies provide information on the binding abilities of the Bcd proteins and the cooperativity of these interactions, ultimately, the only test of a viable interaction of a transcription factor and a target promoter is *in vivo*.

Chapter 7 Transgenic analysis of the *M. domestica tll* promoter

7.1 Introduction

7.1.1 in vivo analysis of the M. domestica tll promoter

In the previous chapters the putative regulatory sequences upstream of the *tll* coding region were described. Thirty-three Bcd binding sites were experimentally identified and the interaction between Bcd and these sites tested by an *in vitro* functional assay. To support the results of the assay it is important to show that these footprinted Bcd binding sites are functional *in vivo*.

The focus of this project is to understand the evolution of the Bcd-*tll* promoter interaction. This interaction involves factors other than Bcd and the binding sites in the *tll* promoter; for example, co-factors which aid activation by Bcd or repressors of Bcd activation such as Dl (Liaw and Lengyel, 1992). It is the combination of all these regulatory factors that results in the wild-type expression patterns of *tll* in *M. domestica* and *D. melanogaster.* Tests of the evolution of the interaction between species should therefore be carried out *in vivo* as it is in this environment that the Bcd-*tll* promoter interaction functions and is subject to selection (Bonneton *et. al.*, 1997).

As the expression of *tll* is generally conserved between the two species it is likely that the *M. domestica cis*-regulatory sequences will drive a similar pattern of expression in *D. melanogaster.* This has been observed with other transgenic constructs between closely related species (Bonneton *et. al.*, 1997; Ludwig *et. al.*, 1998; Wittkopp *et. al.*, 2002). There may be quantitative differences in expression, which are a result of reduced compatibility between the transacting factors and the regulatory sequences (Skaer and Simpson, 2000). There are some differences in the expression of *tll* between *D. melanogaster* and *M. domestica,* such as the difference in the timing of onset of expression in the anterior of embryos. Whether differences such as these are due to changes in *trans* or *cis*-regulatory factors can be determined by comparison of the promoter sequences in a similar genetic background (Wittkopp *et. al.*, 2002).

The study of the regulation *M. domestica* genes is possible by the generation of a transgenic *D. melanogaster* which contains *M. domestica* regulatory sequences (Bonneton *et al.*, 1997; Piccin *et al.*, 2000; Hutson and Bownes, 2003). Since the two species have a conserved early embryonic development interpretation of the transgene expression patterns should be straightforward as it can be assumed that regulation of conserved domains of expression will be by the same factors in both species (Sommer and Tautz, 1991). Unfortunately, the reciprocal test in *M. domestica* of *D. melanogaster tll* promoter sequences would be more problematic, due to the difficulty of transforming *M. domestica* and a lack of mutant lines (White *et al.*, 1996; O'Brochta *et al.*, 1996; Hediger *et al.*, 2001).

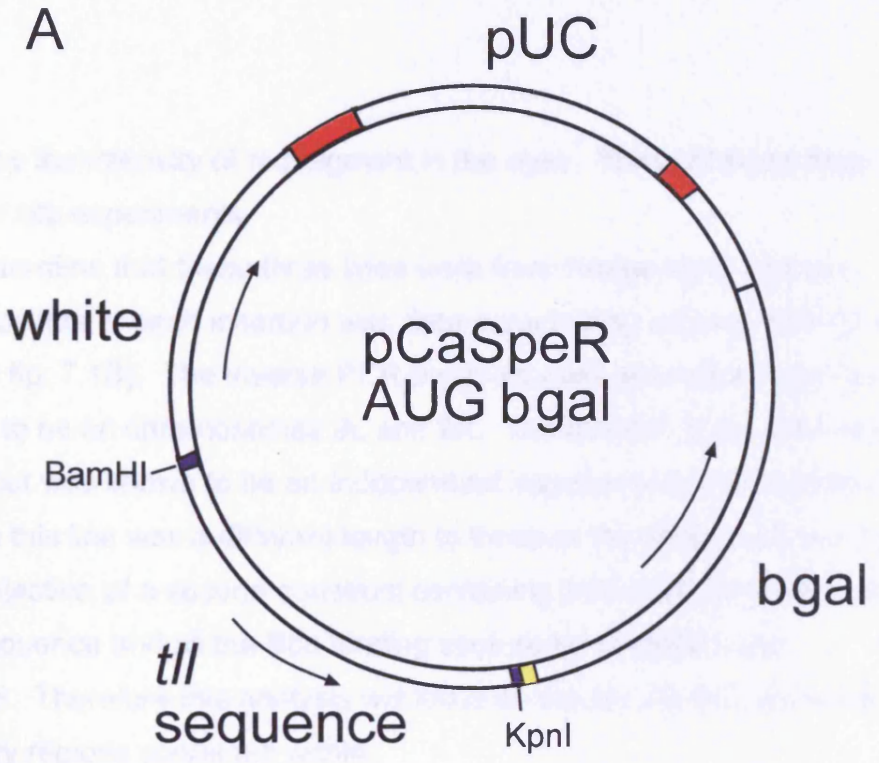
7.1.2 Aims

In order to define the regulatory regions of the *M. domestica tll* promoter, a reporter gene construct containing putative *tll* regulatory sequences was injected into *D. melanogaster*. *In situ* hybridization experiments were carried out to identify the expression pattern driven by the *M. domestica tll* sequences in wild-type *D. melanogaster* and various mutant backgrounds.

7.2 Results

7.2.1 Creating independent transgenic lines of *D. melanogaster* containing the *M. domestica tll* cis-regulatory sequences

The 2.2 kb of sequence immediately upstream of the *M. domestica tll* transcription start site was cloned into the pCaSpeR-AUG-*lacZ* vector. This sequence contains the first 13 Bcd binding sites identified by footprinting (see 4.2.2). The plasmid, M2.*2tll-lacZ*, was injected into *D. melanogaster* and 10 transgenic lines were generated (see 2.2.11 and fig. 7.1A). All the *tll*-lacZ inserts were mapped to chromosomes II or III and had a similar level of expression, as



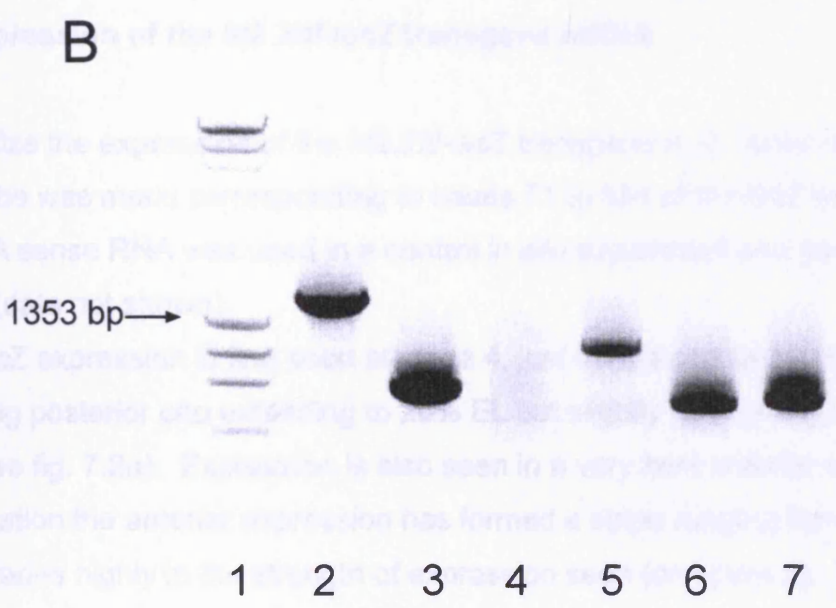


Figure 7.1A. The Casper-AUG-gal plasmid containing 2.2 kb of *M. domestica tll* cis-regulatory sequence (M*tll*). This plasmid contains a *white* and a *lacZ* reporter gene. There is an AUG codon 3' to a mutiple cloning site (Thummel *et al.*, 1988). The 2.2 kb sequence was inserted between the *Bam*HI and *Kpn*I sites of this multiple cloning site.
B. Results of the inverse-PCR of 3 transgenic lines used in the *in situ* hybridization experiments. Lanes 2 and 3 contain reactions with transgenic line 1, lanes 3 and 4 with transgenic line 2 and lanes 4 and 5 with transgenic line 3. The transgenic genomic DNA was digested with *Hin*P1I (lanes 2, 4 & 6) and *Msp*I (lanes 3, 5 & 7 and see 2.2.11). The size markers (and x) are present in lane 1.

determined by the intensity of red pigment in the eyes. Three of these lines were chosen for *in situ* experiments.

To determine that these three lines were from independent insertion events the position of each insertion was determined using inverse PCR (see 2.2.11.4 and fig. 7.1B). The inverse PCR products were sequenced and two were shown to be on chromosomes 3L and 2R. The position of the third was ambiguous but was shown to be an independent insertion since the inverse PCR product from this line was a different length to those of the other two lines (see fig. 7.1B). Injection of a second construct containing 9 kb of *M. domestica tll* upstream sequence and all the Bcd binding sites so far identified was unsuccessful. Therefore this analysis will focus on the M2.2*tll-lacZ* construct and the regulatory regions contained within.

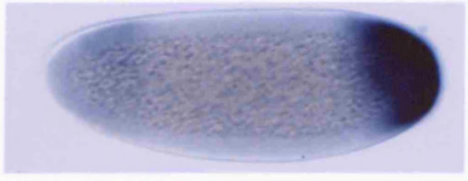
7.2.2 Expression of the M2.2*tll-lacZ* transgene mRNA

To visualize the expression of the M2.2*tll-lacZ* transgene in *D. melanogaster* an RNA probe was made corresponding to bases 71 to 884 of the *lacZ* coding region. A sense RNA was used in a control *in situ* experiment and gave no staining (data not shown).

lacZ expression is first seen at stage 4, just before cellularisation begins, in a strong posterior cap extending to 20% EL but slightly less on the ventral side (n=26, see fig. 7.2a). Expression is also seen in a very faint anterior cap. By cellularisation the anterior expression has formed a stripe ranging from 86-66% EL that varies highly in the strength of expression seen (compare fig. 7.2b & c). The expression of the stripe on the ventral side is absent or very weak.
By early gastrulation (stage 6) the anterior and posterior expression domains begin to move away from the poles and expression on the ventral side is lost (see fig. 7.2d). *lacZ* expression is seen on the ventral side in two stripes either side of the invaginating cells which form the mesoderm (see fig. 7.2d the 2nd column). Later in gastrulation the anterior expression splits down the midline into two separate domains and becomes weaker (see fig. 7.2e). After the completion

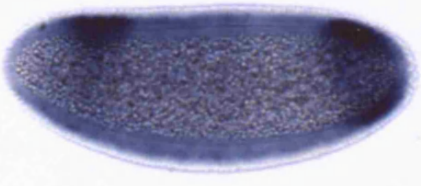
A P Ventral

a

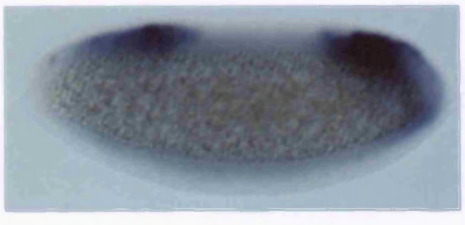


b

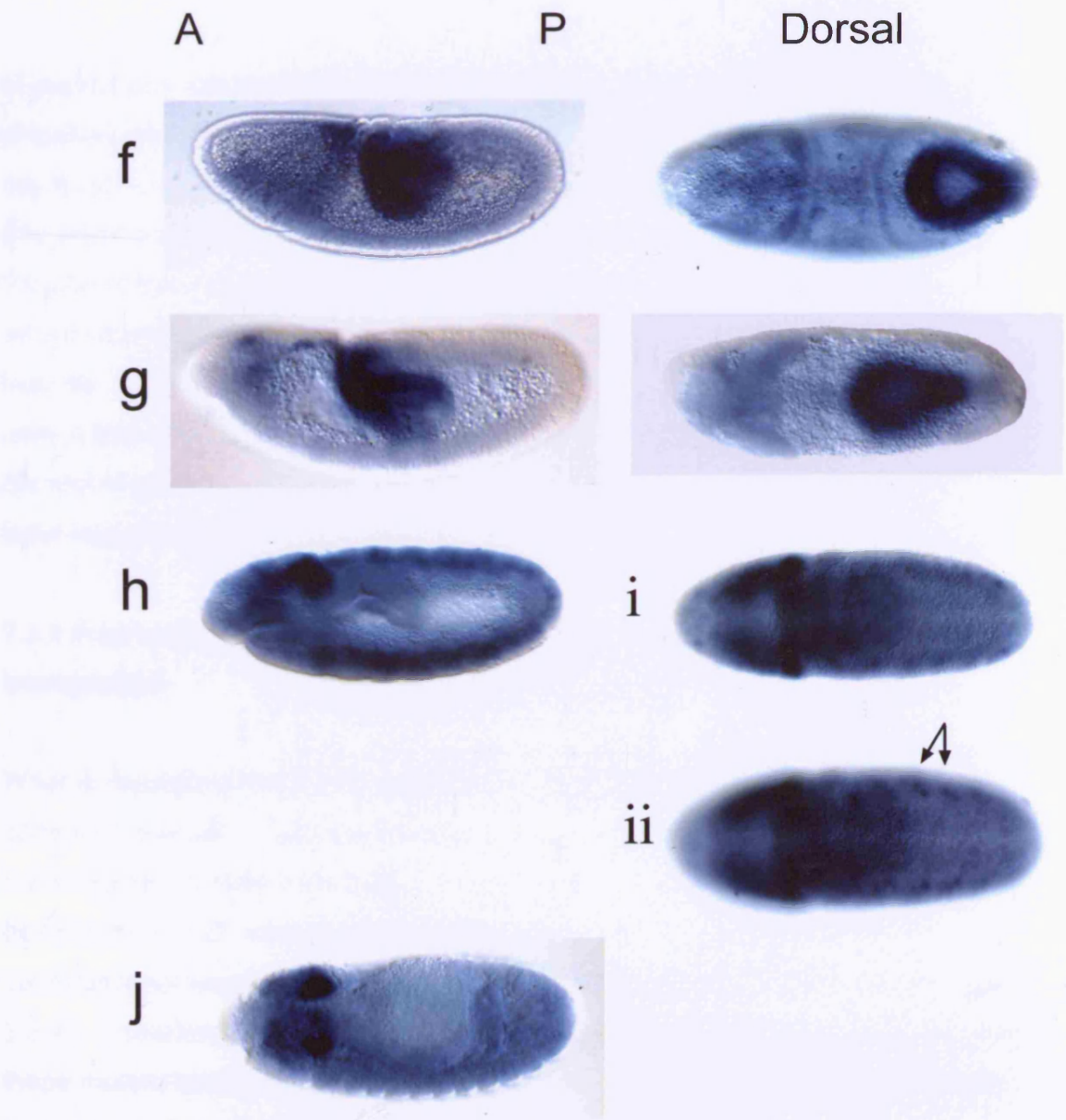
c



d



e



7.2 Expression of the M2.2*tll-lacZ* transgene mRNA in *D. melanogaster* embryos
Embryos on the left are viewed laterally, the anterior (A) to the left and posterior (P) to the right and dorsal to the top. Embryos on the right are viewed ventrally or dorsally, again with anterior to the left. The different stages are as follows:
a. late syncytial blastoderm; b. & c. cellular blastoderm, note embryo c. was stained for longer, hence the greater level of background staining;
d. and e. gastrulation; f. & g. germ band elongation; h. full germ band elongation, i. expression in the developing brain region and ii. expression in the trunk;
j. germ band retraction.

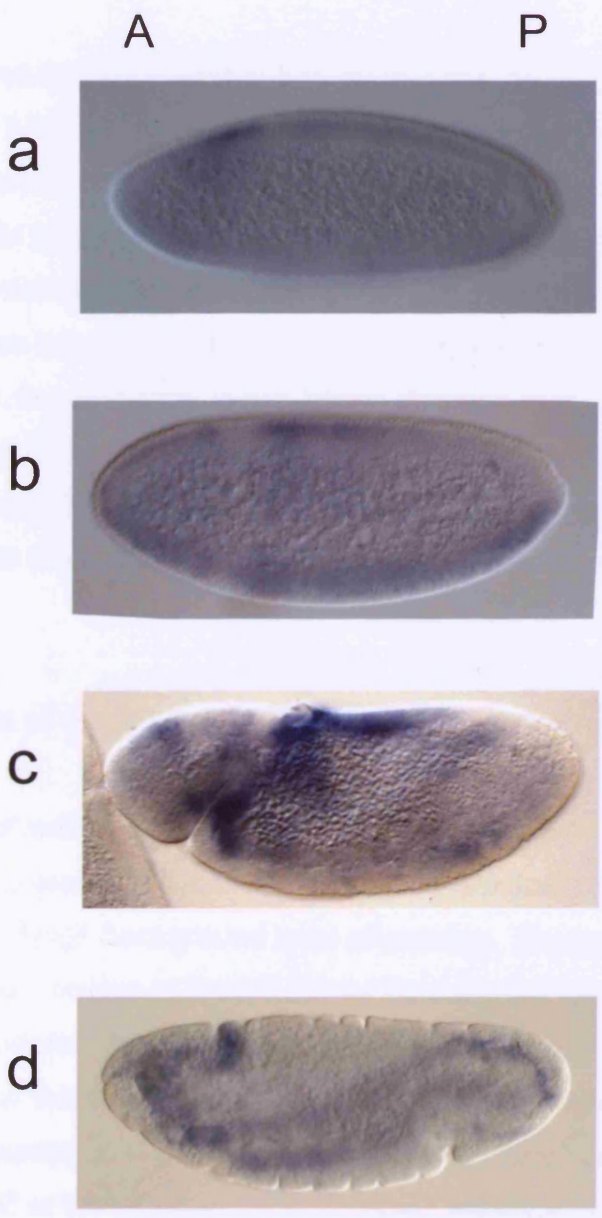
of gastrulation the posterior expression remains strong and continues to move anteriorly with the extending germ band (see fig. 7.2f and g). At this time *lac*Z expression is seen in the invaginating cells that will form the posterior midgut, this expression is restricted mainly to the region of the hindgut (see fig. 7.2g). By the end of germ-band elongation the expression in the hindgut has ceased and strong expression is seen in the anterior, in the region of the developing brain (see fig. 7.2h and hi). At this stage expression is also seen in small groups of cells in the trunk (see fig. 7.2hii). The expression in the trunk has disappeared by the end of germ-band retraction, stage 12, although expression in the developing brain region remains strong (see fig. 7.2j).

7.2.3 Expression of the M2.2*tll-lacZ* transgene in a *D. melanogaster* mutant background

What is regulating the 2.2 kb sequence to give the observed *lac*Z expression patterns observed? Can examination of the *lac*Z expression patterns in *D. melanogaster* mutant lines help in identification of the regulators? As the development of *D. melanogaster* and *M. domestica* is generally conserved the candidates for regulation of the M2.2*tll-lac*Z transgene are Tor, Bcd and Dl (see 3.3.4). Therefore, the expression of the M2.2*tll-lac*Z transgene was examined in these mutant backgrounds. For the schemes of the crosses generated in these experiments see appendix 5.

7.2.4 Expression of the M2.2*tll-lacZ* transgene in maternal *tor*$^{-}$ embryos

In a *tor*$^{-}$ background the expression of *lac*Z was greatly decreased. Expression is first seen in the cellular blastoderm and consists of a faint stripe of expression restricted to the dorsal side (see fig. 7.3a). The position of the stripe corresponds with the position of the stripe seen in M2.2*tll-lac*Z wild-type embryos (see fig. 7.2b and c). Prior to the beginning of gastrulation the expression appears to move posteriorly and faint expression is seen on the ventral side (see fig. 7.3b).

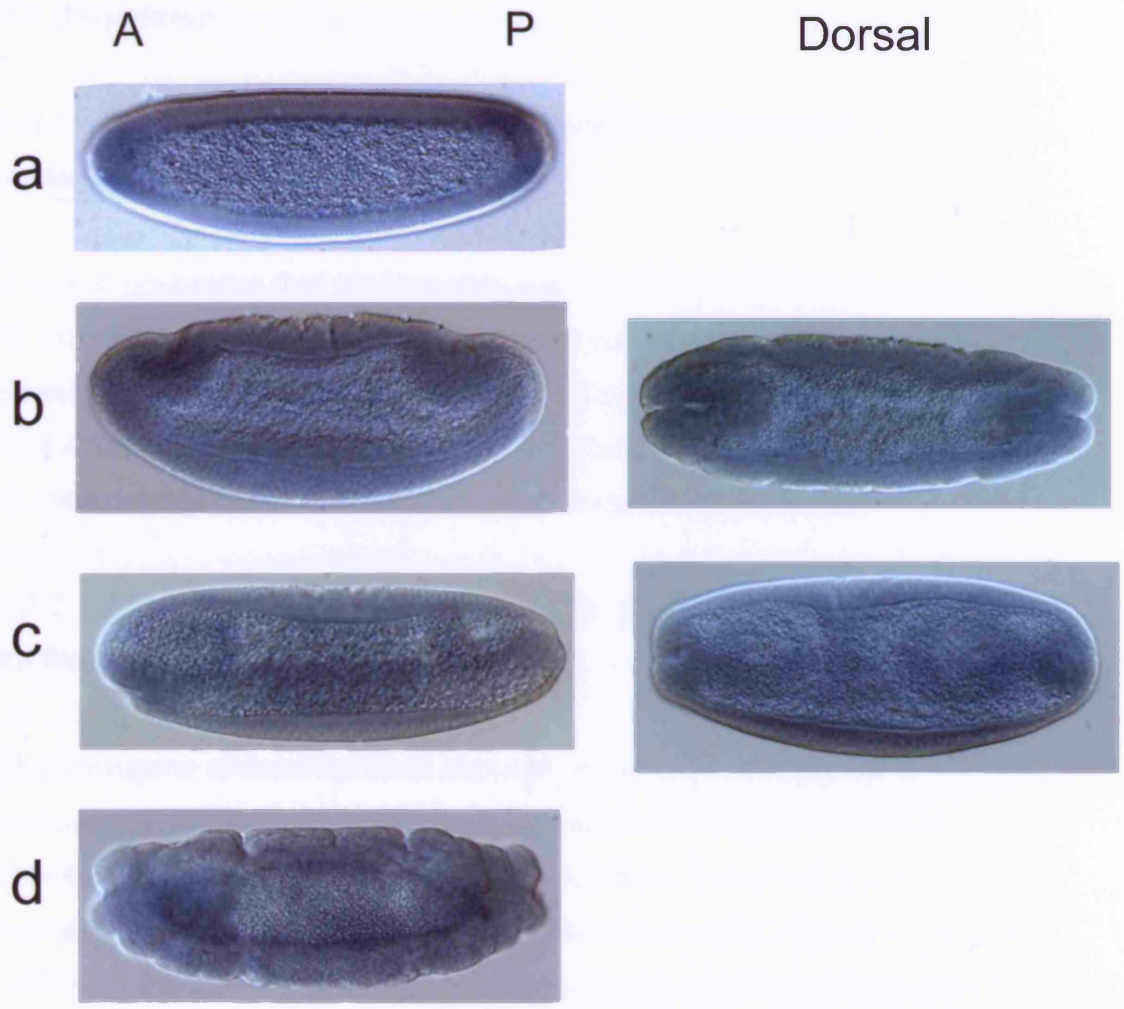


7.3 Expression of the M2.2*tll-lacZ* in maternal *tor*$^{-}$ embryos
Embryos on the left are viewed laterally, the anterior (A) to the left and posterior (P) to the right and dorsal to the top. Embryos on the right are viewed dorsally, again with anterior to the left.
The different stages are as follows: a. & b. cellular blastoderm; c. gastrulation; d. full germ band elongation

The older embryos shown are difficult to stage because of the developmental defects (see fig. 7.3c and d). However, expression is seen in the region anterior to the cephalic furrow and on the dorsal side where the anterior of the extending germ band should be (see fig. 7.3c). In fig. 7.3d the parasegmental furrows have formed in the central part of the embryo. This occurs in wild-type flies in stage 11 and then the germ-band retracts. Expression is seen in the region of the developing brain, in particularly in two lateral domains (see fig. 7.3d). Expression reminiscent of the two stripes of ventral expression is seen throughout the trunk. This expression appears to be on the ventral side in the anterior and then curves up onto the dorsal side, giving the impression of a twist in the embryo (see fig. 7.3d).

7.2.5 Expression of the M2.2*tll-lacZ* transgene in maternal *bcd*$^-$ embryos

The expression of *lacZ* in the *bcd*$^-$ mutant line was very low; therefore, to determine the expression patterns the embryos were stained for a long time, which resulted in a high background level of staining. However, it is still possible to make out the expression of the M2.2*tll-lacZ* transgene, which first appears in the cellular blastoderm. In *bcd*$^-$ mutant embryos two posterior regions develop, one in the place of the missing anterior half of the embryo (Berleth *et al.*, 1988). Indeed the first expression pattern is in two terminal caps, resembling the expression of *lacZ* at the posterior of M2.2*tll-lacZ* wild-type embryos (see fig. 7.4a). The low level of expression makes it difficult to comment further on the differentiation of these caps at this stage. During gastrulation the expression domains move dorsally with the extending germ-band and there appears to be expression along the ventral side of the embryo (see fig. 7.4b and c). This expression on the ventral side appears to last until stage 11 when the parasegmental furrows are formed and then disappears along with the expression on the dorsal side (see fig. 7.4d).



7.4 Expression of the M2.2*tll-lacZ* in maternal *bcd* $^{-}$ embryos
Embryos on the left are viewed laterally, the anterior (A) to the left and posterior (P) to the right and dorsal to the top. Embryos on the right are viewed dorsally, again with anterior to the left. The different stages are as follows: a. early cellular blastoderm; b. & c. gastrulation; d. full germ band elongation

7.3 Discussion

7.3.1 Expression of a *M. domestica* cis-regulatory element in *D. melanogaster*

A region containing Bcd binding sites was identified in the region upstream of the *M. domestica tll* coding region. To test the function of this sequence it was transformed into *D. melanogaster* attached to a *lacZ* reporter. *In situ* hybridization experiments revealed the pattern of expression of the transgene in *D. melanogaster* wild-type, *tor*$^-$ and *bcd*$^-$ mutant backgrounds.

7.3.2 Evidence for sequences responsive to the Tor, Bcd and Dl in the 2.2 kb fragment

The transgene allowed for an *in vivo* comparison of the regulation of *tll* in *D. melanogaster* and *M. domestica*. As regulation is thought to be conserved between *D. melanogaster* and *M. domestica* there should be evidence of regulation by Tor, Bcd and Dl (Sommer and Tautz, 1991; Liaw and Lengyel, 1992). Is there evidence for these three separate regulatory interactions?

Expression of *lacZ* mRNA in a posterior cap indicates that the elements responsive to activation by Tor are found within the 2.2 kb cis-regulatory sequences (see fig. 7.2a, b and c). The retraction of this expression from the anterior tip of the embryo indicates that the sequences contain elements responsive to Tor mediated repression of *tll* (see fig. 7.2b and c).

Evidence for regulation by Bcd of the 2.2 kb fragment comes from the appearance of the stripe at cellular blastoderm which is activated by Bcd in *D. melanogaster*. The repression of *tll* at the anterior tip at this stage is also due to Bcd in *D. melanogaster* and so is likely to involve Bcd in *M. domestica*. However, the expression of the stripe in M2.2*tll-lacZ* embryos is weaker than that seen in *M. domestica* and suggests that some of the sequences responsive to activation by Bcd are missing from the transgene (see fig. 7.2b and c).

The anterior stripe is absent on the ventral side of the embryo and this indicates that Dl repression of the stripe in *M. domestica* also maps to the 2.2 kb of sequence present in the transgene. This suggests the putative Dl binding sites seen in the 2.2 kb sequence are functional (see fig. 4.18).

7.3.3 Bcd and Tor regulate the M2.2*tll-lacZ* transgene in *D. melanogaster*

The expression patterns indicate that Tor, Bcd and Dl regulatory elements are present in the M2.2*tll-lacZ* transgene. If this is correct then the expression of *lacZ* should be altered in embryos mutant for these transcription regulators. Indeed in a *tor*$^{-}$ mutant the expression of the early posterior cap is lost (see fig. 7.3a) but in *bcd*$^{-}$ mutant embryos this posterior expression is present (see fig. 7.4a). These results confirm that the posterior domain is activated by Tor and that at least one Tor-RE is present in the 2.2 kb sequence. Further evidence of regulation by Tor is suggested by the faint staining at the anterior of *tor*$^{-}$ mutant embryos during the cellular blastoderm (see fig. 7.3b). This could represent loss of the Tor regulated repression of *tll,* which prevents expression at the anterior tip of M2.2*tll-lacZ* wild type embryos. This derepression indicates that elements responsive to repression as well as activation by Tor are present within the 2.2 kb sequence (see fig. 4.18).

The anterior stripe is present in *tor*$^{-}$ mutant embryos, this shows that it is being activated by a factor other than TorRTK, and the likely candidate is Bcd (see fig. 7.2a). However regulation of the stripe by Bcd cannot be determined in a *bcd*$^{-}$ mutant because in the absence of Bcd the anterior is not patterned properly (Berleth *et al.*, 1988). To confirm that Bcd was activating the stripe in *D. melanogaster* the position of the stripe was compared in embryos with one, two or four copies of maternal *bcd* (Pignoni *et al.*, 1992). As expected for a Bcd activated domain of expression the stripe was shifted to the anterior and posterior in embryos with one and four copies of maternal Bcd respectively, in comparison to wild-type. To confirm Bcd activation of the anterior stripe it would

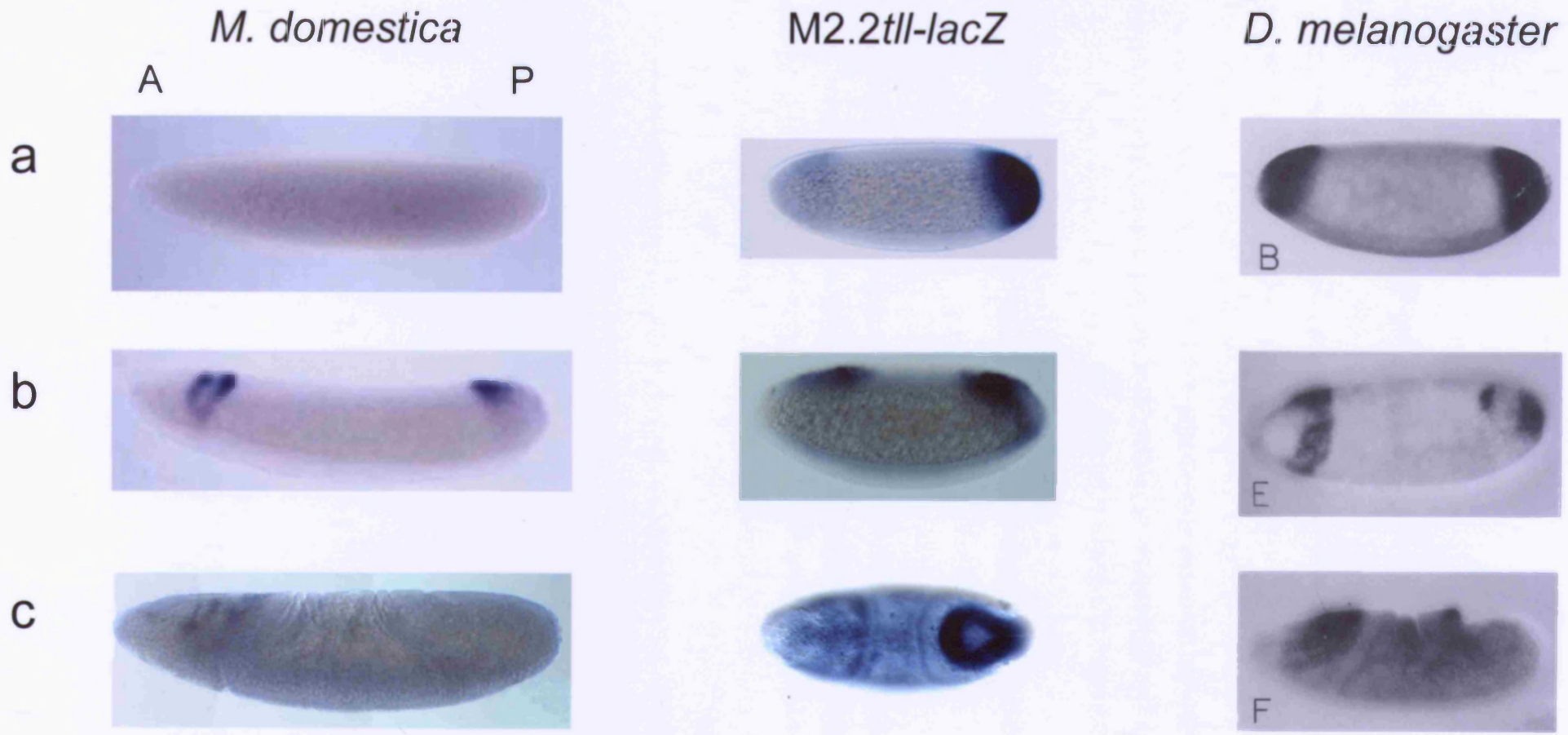
be necessary to test the expression of the transgene in a background of one or four copies of maternal *bcd*.

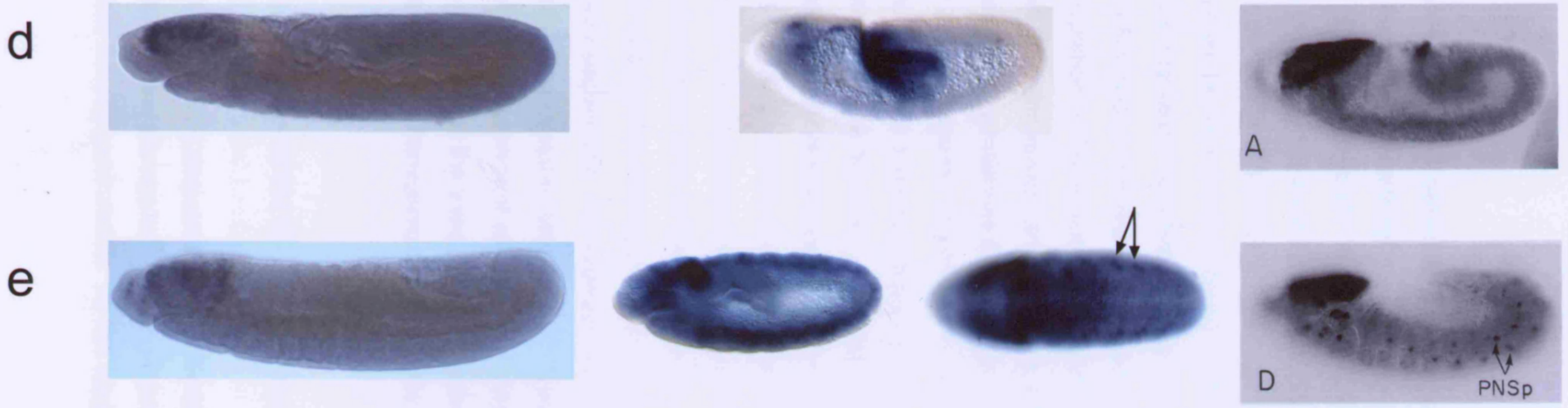
The expression of the anterior stripe is observed to be weaker in *tor*$^{-}$ mutant embryos than in wild type embryos. In *D. melanogaster*, TorRTK activity enhances the activation by Bcd of the stripe (Liaw and Lengyel, 1992; Janody *et al.*, 2001). Therefore the weak expression of the stripe in *tor*$^{-}$ embryos could be a combination of an incomplete regulatory module responsive to Bcd and loss of TorRTK enhancement of activation by Bcd.

7.3.4 Comparison of expression patterns between the M2.2*tll-lacZ* transgene and *M. domestica tll*

The 2.2 kb sequence of the *M. domestica tll* promoter contains much of the regulatory sequences needed to drive correct expression of the *tll* gene. Therefore although incomplete the 2.2 kb cis-regulatory sequence may be used to examine the evolution of regulatory differences between *D. melanogaster* and *M. domestica tll*.

The expression of the M2.2*tll-lacZ* transgene resembles that of *M. domestica tll* in many ways. However some expression domains are more reminiscent of *D. melanogaster* and some appear to be ectopic. These different expression patterns and the potential evolutionary implications will be discussed below. Importantly, regulation of expression in a posterior cap and anterior stripe are conserved in both species and the transgene. This is in accord with the necessity of these expression domains for the correct *tll*-dependent development of the embryo in *D. melanogaster* (Pignoni *et al.*, 1990; see fig. 3.6b and 7.2b). Expression of *tll* in the developing head is conserved between *D. melanogaster* and *M. domestica* and this is reflected by a similar expression of the transgene (see fig. 7.5c, d and e). It is likely the factors that regulate *tll* expression in the head are conserved between the species. A more detailed study, for example by identification of the tissues in which *tll* is expressed, is necessary for further comparison of these later stages of *tll* regulation.


M. domestica
M2.2tll-lacZ
D. melanogaster
A
P
a
b
c
B
E
F



7.5 Comparing the expression patterns of *tll* in *M. domestica* and *D. melanogaster* and of the M2.2*tll-lacZ* transgene
Embryos are viewed laterally, the anterior (A) to the left and posterior (P) to the right and dorsal to the top, except for c. M2.2*tll-lacZ* which is a dorsal view. The different stages are as follows: a. syncytial blastoderm; b. cellular blastoderm; c. gastrulation; d. germ band elongation; e. germ band retraction, the arrows indicate expression in the small groups of cells in the trunk. The *D. melanogaster* images are taken from Pignoni *et al.*, 1990.

In *D. melanogaster* the terminal system activates expression of genes, such as *tll*, *hb* and *otd*, in an anterior cap in the syncytial blastoderm (Finkelstein and Perrimon, 1990; Ronchi *et al.*, 1993). This activation is missing in *M. domestica* and is attributed to a difference in regulation by Tor between the species. The appearance of a faint cap in the anterior of M2.2*tll-lacZ* syncytial embryos resembles the expression of *tll* in *D. melanogaster* (see fig. 7.5a). Therefore, the cis-regulatory sequence retains the elements necessary to produce a *D. melanogaster*-like expression pattern. This suggests that the difference in regulation by Tor is due to differences in the transacting regulatory factors between *D. melanogaster* and *M. domestica* (Wittkopp *et al.*, 2002).

There are further similarities between the expression of the M2.2*tll-lacZ* transgene and *D. melanogaster tll*, which differ from *M. domestica*. Firstly the M2.2*tll-lacZ* anterior stripe does not differentiate into two stripes along the anterior-posterior axis (see fig. 7.5b). Secondly, during germ-band retraction there is expression in discrete groups of cells in the trunk, which is absent in *M. domestica* (see fig. 7.5e). As the promoter sequence is driving a wild-type expression pattern in *D. melanogaster* the differences between *D. melanogaster* and *M. domestica* appear to be in trans. However, the differences between the *M. domestica tll* wild-type and transgene expression patterns could be due to missing regulatory sequences. For example, expression of *tll* in groups of cells in the trunk could be the ancestral state, which has subsequently been lost in *M. domestica* due to the presence of an inhibitory element located beyond the 2.2 kb sequences contained in the transgene. To answer these questions it would be necessary to create a transgenic containing the entire *M. domestica tll* promoter.

7.3.5 Ectopic expression of the M2.2*tll-lacZ* transgene

Introducing a cis-regulatory sequence into a different species can result in ectopic expression of the transgene. Ectopic expression of the M2.2*tll-lacZ* transgene is seen in early gastrulation. At this stage cells on the ventral side of

the embryo invaginate to form the ventral furrow. These cells will go on to form the mesoderm. *lacZ* expression is seen in two stripes, which mark the edges of the ventral furrow and resemble expression of *snail* (*sna*) at this stage (see fig. 7.2d, the 2nd column). The expression of *sna* is activated directly by Dl and Twist (Ip *et al.*, 1992) and Dl is a potential regulator of these ectopic stripes of expression. Indeed, there is evidence for a Dl responsive region in the transgene because of the repression of the anterior stripe on the ventral side of transgenic embryos (see 7.3.3).

Dl is a transcriptional activator, but in the presence of the transcription factor Dri and the co-repressor Gro it is converted into a repressor (Dubnicoff *et al.*, 1997; Valentine *et al.*, 1998). This mechanism appears to be conserved between *D. melanogaster* and *M. domestica* to the extent that the *tll* anterior stripe is repressed on the ventral side of the embryo. However the mechanism may have diverged such that in *D. melanogaster* transgenic embryos Dl is activating expression from the *M. domestica* sequences on the ventral side. This change could involve other activating factors present at gastrulation or in the cells in which the ectopic activation is observed, such as the transcription factor Twist, which is present in the ventral part of the embryo (Thisse *et al.*, 1988). The ectopic expression may also be attributed to incomplete cis-regulatory regions present in the transgene. To discover if this expression is due to Dl activation it would be necessary to study the expression of the transgene in a *dl* mutant background.

The expression of the transgene in the posterior is much stronger and is present for longer than both wild-type *tll* expression patterns (see fig. 7.5d). Interestingly, ventrally expressed genes such as *twist* and *sna* are also expressed in the extending germ band (Thisse *et al.*, 1988; Ip *et al.*, 1992). Therefore, it is possible the activating factors responsible for the ventral expression are enhancing this expression in the posterior.

7.3.6 Summary

The expression patterns of the transgene show that the regulation of *M. domestica tll* is generally conserved with *D. melanogaster tll*. There is evidence for both activation and repression of the transgene by Bcd, as predicted by the *DNase*I footprinting experiments. However, the weak expression of the Bcd dependent stripe suggests further Bcd activating sites are absent from the transgene. Such sites could be those present in region B of the *M. domestica* sequence (fig. 4.16). It is also possible that the weak activation of the stripe is due to the divergence of other transacting factors between *M. domestica* and *D. melanogaster*. Indeed it has been shown that there is a difference between the species in the regulation of *tll* and other genes by the terminal system (see 7.3.4).

Excluding the differences in the regulation of some secondary expression patterns it is remarkable that both of these promoter sequences can drive a similar and complex expression pattern of *tll* within *D. melanogaster* even though they are unalignable. The implications of these findings and the differences in *tll* regulation between the species will be discussed further in chapter 9.

Chapter 8 Characterisation of the *cad* gene in *M. domestica* and the interaction between *cad* and Bicoid in this species

8.1 Introduction

8.1.1 Why study the Bcd *cad* mRNA interaction?

To determine the molecular processes by which a conserved functional interaction evolves, the Bcd protein interaction with the *hb* promoter was compared between *D. melanogaster and M. domestica.* It was shown that there were differences in the Bcd DNA binding domain, other functional domains and the structure of the *hb* promoter between the two species (Bonneton *et. al.*, 1997). The Bcd-*hb* promoter interaction is part of a network of Bcd interactions, which includes many other target gene promoters and regulatory factors. Therefore, to understand the cause of differences observed between the Bcd-*hb* interaction in these two species, the analysis was extended to other genes transcriptionally regulated by Bcd (Shaw *et. al.*, 2002; McGregor, 2002).

The Bcd protein is unusual in that it is able to bind both DNA and RNA, both functions being mediated by the homeodomain (Dubnau and Struhl, 1996; Rivera-Pomar *et. al.*, 1996). In *D. melanogaster,* Bcd binds the 3' UTR of the *cad* mRNA and represses translation of Cad via a cap dependent mechanism (Dubnau and Struhl, 1996; Niessing *et. al.*, 1999; see 1.15). The PEST domain of Bcd is necessary for this function (Rivera-Pomar *et. al.*, 1996). The Bcd-*cad* mRNA interaction may have influenced the evolution of the Bcd protein sequence. The recent origin of Bcd and the absence of other RNA binding homeodomains, suggests that this RNA binding function is a recently evolved mechanism. Therefore, selection for the RNA binding ability could have impacted on the evolution of the Bcd protein, *cad* mRNA and other Bcd-dependent gene promoters.

8.1.2 The role of *cad* in posterior determination

cad is a homeobox containing transcription factor which is involved in determining the development of the posterior region of the embryo. In *D.*

melanogaster, Cad is necessary for the correct development of the hindgut, the anal pads, the malphigian tubules and the eighth abdominal segment (Wu and Lengyel 1998). The presence of ectopic Cad in the anterior of the embryo disrupts head development and this is probably due to ectopic activation of *cad* target genes in the anterior (Macdonald and Struhl, 1986; Niessing *et. al.*, 1999). The role of *cad* genes as a determinant of posterior structures early in embryonic development is highly conserved in metazoans (Marom *et. al.*, 1997).

8.1.3 *D. melanogaster cad* mRNA and protein expression patterns

In *D. melanogaster* a maternal *cad* transcript is homogeneously distributed throughout the embryo before fertilisation (Macdonald and Struhl, 1986). Repression of this maternal *cad* mRNA translation by Bcd results in a gradient of Cad protein along the anterior-posterior axis, with the peak of expression at the posterior (Macdonald and Struhl, 1986). Subsequently, both protein and transcript are degraded in a posterior direction and by the end of cellularisation have virtually disappeared (Mlodzik and Gehring, 1987). The zygotic transcript of *cad* appears at this stage in a stripe three to four cells wide from 13-19% EL (Mlodzik and Gehring, 1987). Tailless, Brachyenteron and Kruppel regulate zygotic cad expression and the anterior border of the zygotic stripe is defined by Hunchback repression (Singer *et. al.*, 1996; Liu and Jack, 1992; Schultz and Tautz, 1995). This stripe moves dorsally with the extending germ-band and is expressed in the anlagen for the terminal abdominal structures and the hindgut. There is expression of *cad* mRNA in the invaginating posterior midgut and malphigian tubules. During germ-band retraction Cad is present in the hind-gut, posterior mid-gut, malphigian tubules and at the posterior tip, in the cells which will form the anal pad; this expression persists in the larvae (Macdonald and Struhl, 1986).

8.1.4 Aims

To clone and sequence the *cad* gene in *M. domestica* and identify the expression pattern of the gene. Comparison of these results with *D. melanogaster* will indicate if the role of *cad* is conserved between the species. Subsequently to confirm the interaction of Bcd and *cad* mRNA in *M. domestica* and compare with *D. melanogaster*.

8.2 Results and discussion

8.2.1 Sequencing of the *M. domestica cad* gene

To clone the *cad* gene from *M. domestica* degenerate *cad* homeodomain primers were designed and used in PCR reactions with *M. domestica* genomic DNA (see fig. 8.1 and 2.1.4). The resulting fragment was predicted to encode a peptide of 54 amino acids, which differed in only one residue from the *D. melanogaster cad* homeodomain (Mlodzik *et. al.*, 1985). Further primers were used in sPCR and 3' RACE experiments to sequence the C-terminal end of the coding region and the 3' UTR (see fig. 8.1, 2.2.4 and 2.2.7). PCR of genomic DNA confirmed the sequence of the 3' UTR (see fig. 8.1). The 3' UTR is approximately three times longer than that of the *D. melanogaster cad* mRNA (see fig. 8.2). Only one polyA signal was observed in the *M. domestica cad* mRNA, 20 bp upstream of the polyA tail. In *D. melanogaster* there are two polyA signals corresponding to the maternal and zygotic transcripts (see fig. 8.2). In both *B. mori* and *T. castaneum* only one polyA signal is seen even though in *T. castaneum* there is both a maternal and zygotic transcript (Xu *et. al.*, 1994; Schulz *et. al.*, 1998).

The 5' end of the *cad* gene was sequenced with a combination of sPCR and degenerate RT-PCR to avoid sequencing the entire intron, which is 10.5 kb in *D. melanogaster* (see fig. 8.1 and 2.1.4). The position of the intron just 5' of the homeodomain was the same in *D. melanogaster* and *M. domestica* (see fig. 8.2). This intron position is highly conserved amongst vertebrate and invertebrate

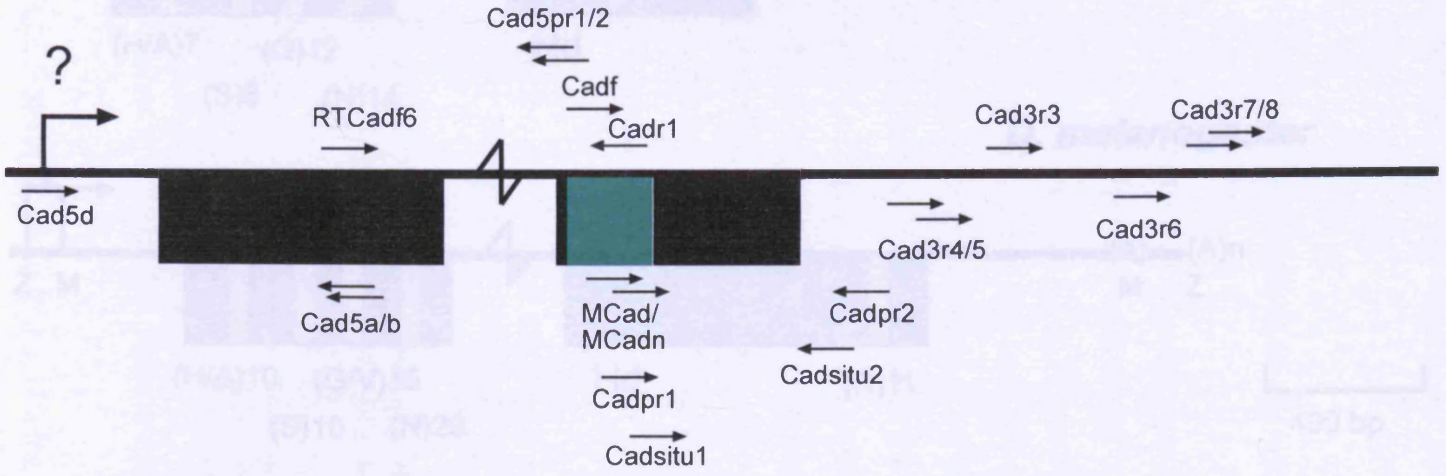


Figure 8.1 Sequencing of *M. domestica cad*.
Primer position and orientation are shown in relation to the *cad* transcript – see text for their use. The arrow indicates the transcription start site and the black boxes the coding region, the green box represents the homeodomain. The length of the intron is unknown.

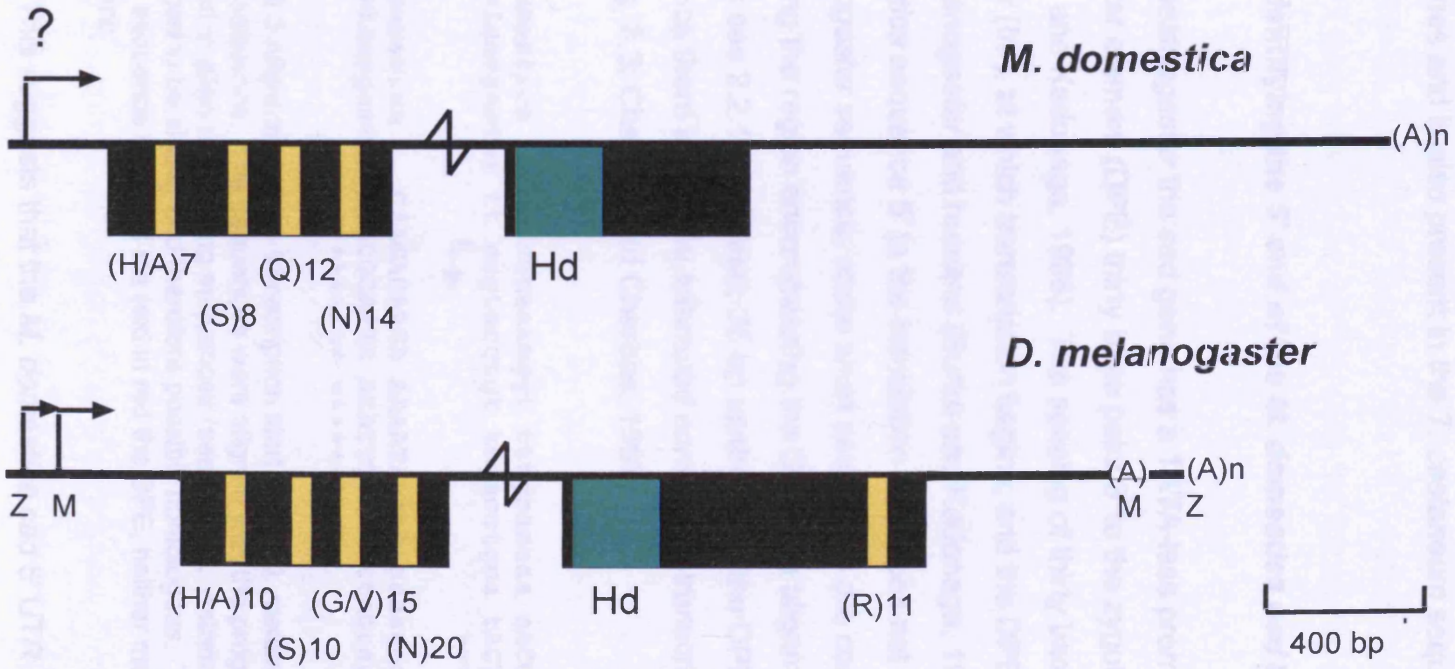


Figure 8.2 Comparison of the *M. domestica cad* transcript structure with that of *D. melanogaster.*
The arrows indicate the transcription start sites, z = zygotic and m = maternal and the question mark indicates that the transcription start is unknown in *M. domestica*. The boxes represent the coding regions; green for the homeodomain. The yellow boxes represent homopolymeric runs of amino acids, which amino acid and number of residues present is given below each box. The intron is marked by a zig-zag line and is 10.5 kb in *D. melanogaster*. The polyA signals are marked with an 'A' (M – maternal; Z – zygotic).

cad genes and is also present in the *T. castaneum* sequence (Schulz *et. al.*, 1998).

8.2.2 Identifying the 5' end of the *M. domestica cad* transcript

In *D. melanogaster* the *cad* gene has a TATA-less promoter and a downstream promoter element (DPE) thirty base pairs 3' to the zygotic transcription start site (Burke and Kadonaga, 1996). The spacing of thirty base pairs between the Initiator (Inr), at which transcription begins, and the DPE is conserved between *D. melanogaster* and humans (Burke and Kadonaga, 1997). Although the *M. domestica* sequence 5' to the translation start does not align well with the *D. melanogaster* sequence, some short sequences are conserved between the two including the region encompassing the DPE (the alignment was done with DiAlign see 2.2.12). Indeed, 35 bp upstream of the DPE in the *M. domestica* sequence there is a loose arthropod consensus transcription start site of GCATT (see fig. 8.3; Cherbas and Cherbas, 1993).

```
M. domestica     gc attaaaaggt tttgcaacaa aACCAAGACA ACAACGCGTG
D. melanogaster  tt cagtacgtgt tcgacctgca tACTGAGTTG ACGTCGCCAC
                                          ********* **********

M. domestica     CAAGACAAGA AAAATAAAAA aaaagacgaa agggag---A
D. melanogaster  GCACGGAATA ACATTTCAAA ttagtagatt agccctgcaA
                 ********** **********                    *
```

Figure 8.3 Alignment of the transcription start site of *D. melanogaster* with *M. domestica* 5' UTR sequence. The sequences were aligned with the programme Di Align that was designed to align non-coding sequences (see 2.2.12). Asterisks indicate sequences that are judged to be similar and therefore possibly homologous. The text in blue is the Initiator sequence (Inr) and the text in red the DPE, neither motif has been verified in *M. domestica*.

This suggests that the *M. domestica cad* 5' UTR is 321 bp long resulting in an overall transcript length of approximately 3.3 kb. This compares to the *D. melanogaster* zygotic and maternal transcripts of 2.6 and 2.4 kb respectively. The expression data for *M. domestica cad* mRNA suggest the presence of both a maternal and zygotic transcript (see 8.2.5). Unfortunately a Northern blot to

determine the full length of the mRNA and number of transcripts was unsuccessful. Apart from the much greater length of the *M. domestica cad* 3' UTR, the general structure of the *cad* transcript are similar between *M. domestica* and *D. melanogaster* (see fig. 8.2).

8.2.3 Comparison of the *M. domestica* Cad protein with other Cad homologues

The putative *M. domestica* Cad protein is 401 amino acids in length and has 60% identity to *D. melanogaster* Cad, which is 472 amino acids in length (see fig. 8.4; Mlodzik and Gehring, 1987). The homeodomain has three differences in comparison to *D. melanogaster* Cad (Niessing *et. al.*, 2000). The regions flanking the homeodomain are very similar and are highly conserved between vertebrate and invertebrate Cad sequences (Marom *et. al.*, 1997; see fig. 8.4). Interestingly, a tyrosine is present at position 24 of helix one of the homeodomain of both *M. domestica* and *D. melanogaster* Cad, whilst in all other *cad* genes and homeobox proteins this residue is a phenylalanine (see fig. 8.4; Mlodzik and Gehring, 1987). The relevance of this change is unknown but the high conservation of phenylalanine amongst other Cad and homeodomain proteins suggests it is functionally relevant.

Apart from the homeodomain one other functional domain has been identified in the Cad protein. This is a conserved hexapeptide motif, Y/C EWMR K/R, N-terminal to the homeodomain (see fig. 8.4). This motif is found at this same position in many homeodomain proteins and is involved in the interaction with Pbx proteins (Marom *et. al.*, 1997). The interaction with Pbx proteins can alter the sequence recognition of the Hox proteins. The hexapeptide motif is present in both *M. domestica* and *D. melanogaster*, although the critical tyrosine residue is replaced by a phenylalanine in *D. melanogaster* (see fig. 8.4).

Both *M. domestica* and *D. melanogaster* Cad proteins have homopolymeric repeats, some of which are shared (see fig. 8.2 and 8.4). These runs of amino acids in the Cad proteins vary in length between the species and in

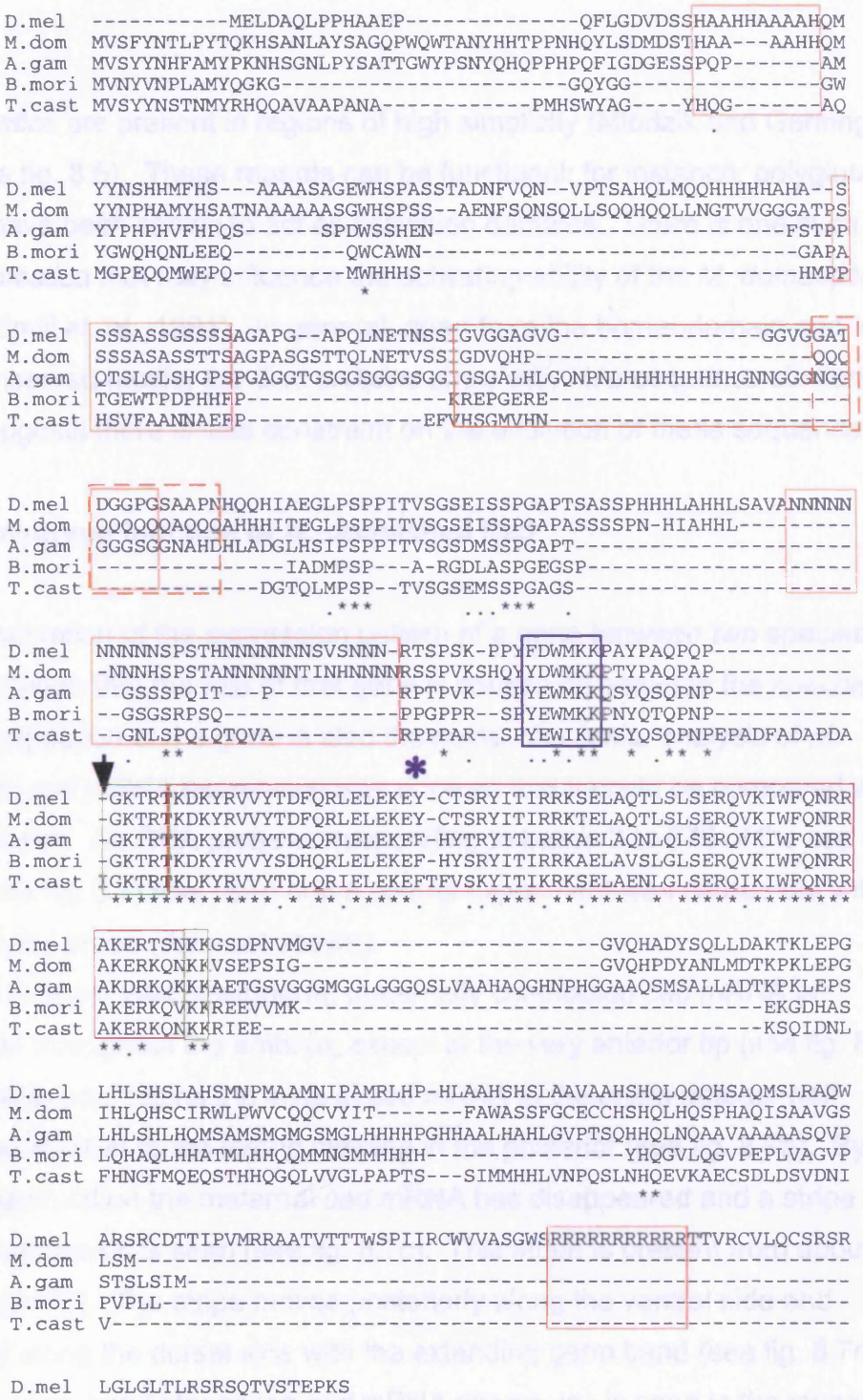


Figure 8.4 Cad protein comparison between insect species:
D. mel - *Drosophila melanogaster* (Mlodzik and Gehring, 1987),
M. dom - *Musca domestica* (this work), A.gamb – *Anopheles gambiae* (Devenport,M.P. and Eggleston,P., AF119382), B.mori – *Bombyx mori* (Xu *et. al.*, 1994), T. cast - *Tribolium castaneum* (Schultz *et. al.*, 1998). The homeodomain is boxed in red, the novel Tyrosine indicated by the blue asterix. The conserved intron position is indicated by the blue arrow head. The highly conserved sequences flanking the homeodomain are shown by green boxes. The conserved hexapeptide is marked by a blue box. The homopolymeric runs are indicated by the orange boxes.

M. domestica are present in regions of high simplicity (Mlodzik and Gehring, 1987; see fig. 8.5). These repeats can be functional; for instance, polyglutamine repeats have been shown to act as activation domains. There is one such repeat in *M. domestica* that may influence the activating ability of the *M. domestica* Cad protein (Emili *et. al.*, 1994). In general, apart from the homeodomain and some small conserved motifs, the Cad proteins show very little sequence similarity, which suggests there is little constraint on the evolution of these sequences.

8.2.4 Conserved function of *M. domestica* Cad

The conservation of the expression pattern of a gene between two species is a good indication that the role of that gene is conserved between the species and that the regulation of the gene is also the same. An *in situ* analysis of *M. domestica cad* mRNA expression was done so that it could be compared with *D. melanogaster.* An RNA probe corresponding to bases 3 to 675 of the *cad* coding region (see fig. 8.1) was used and a control experiment was carried out with a sense strand probe (data not shown).

In the syncytial blastoderm, maternally contributed *cad* mRNA is distributed throughout the embryo, except at the very anterior tip (see fig. 8.7a). As cellularisation begins the level of *cad* mRNA in the entire anterior half decreases relative to the strong staining in the posterior (see fig. 8.7b). By the start of gastrulation the maternal *cad* mRNA has disappeared and a stripe of zygotic expression is seen (see fig. 8.7c). This stripe is present from about 15 to 21% EL (n=15). The stripe moves posteriorly along the ventral side and anteriorly along the dorsal side with the extending germ band (see fig. 8.7d). By maximal germ-band elongation *cad* mRNA expression is seen in the invaginating hindgut and posterior midgut (see fig. 8.7e). There is also expression in the eighth abdominal segment, which is clearly seen from the dorsal side of the embryo (see fig. 8.7e). After germ band retraction during the final stages of embryogenesis expression of *cad* mRNA is seen in the internalized hindgut, posterior midgut and malphigian tubules and also at the posterior tip where the

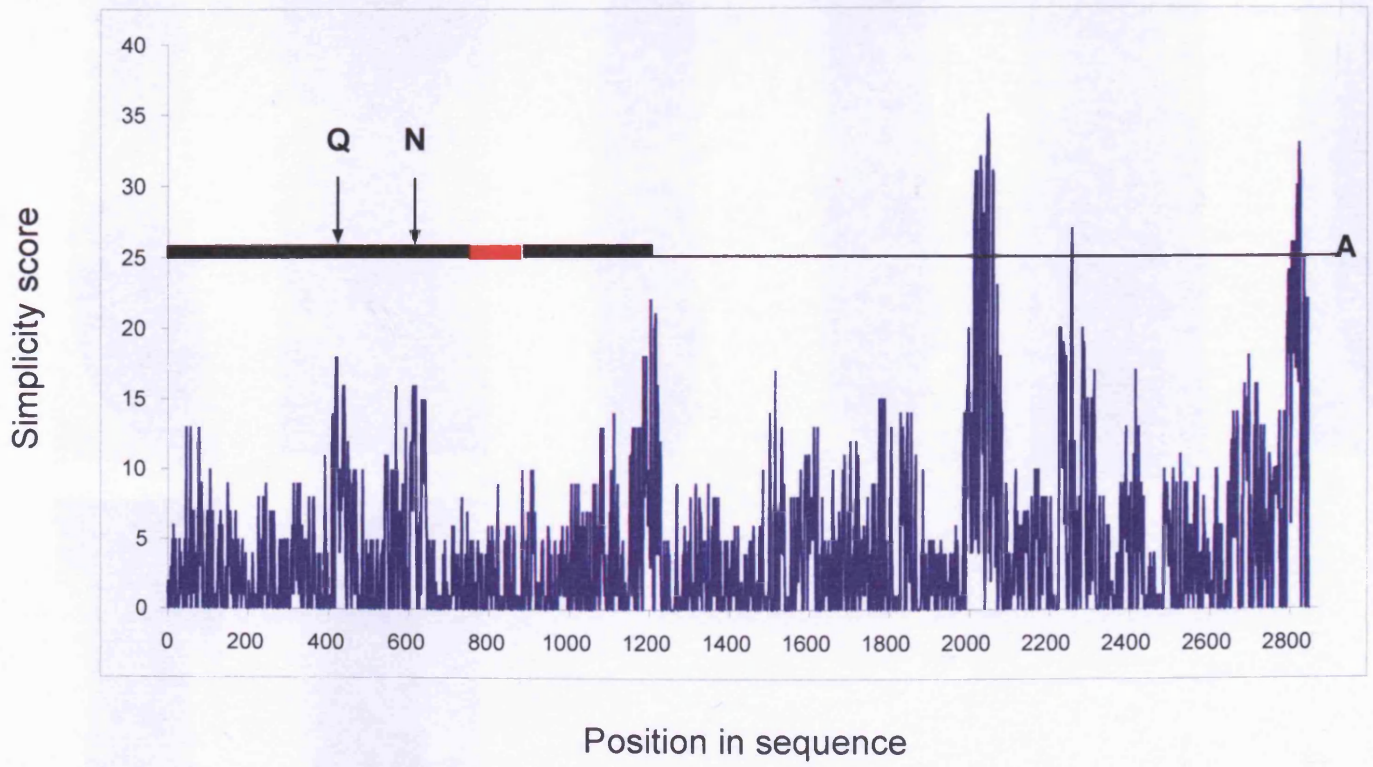
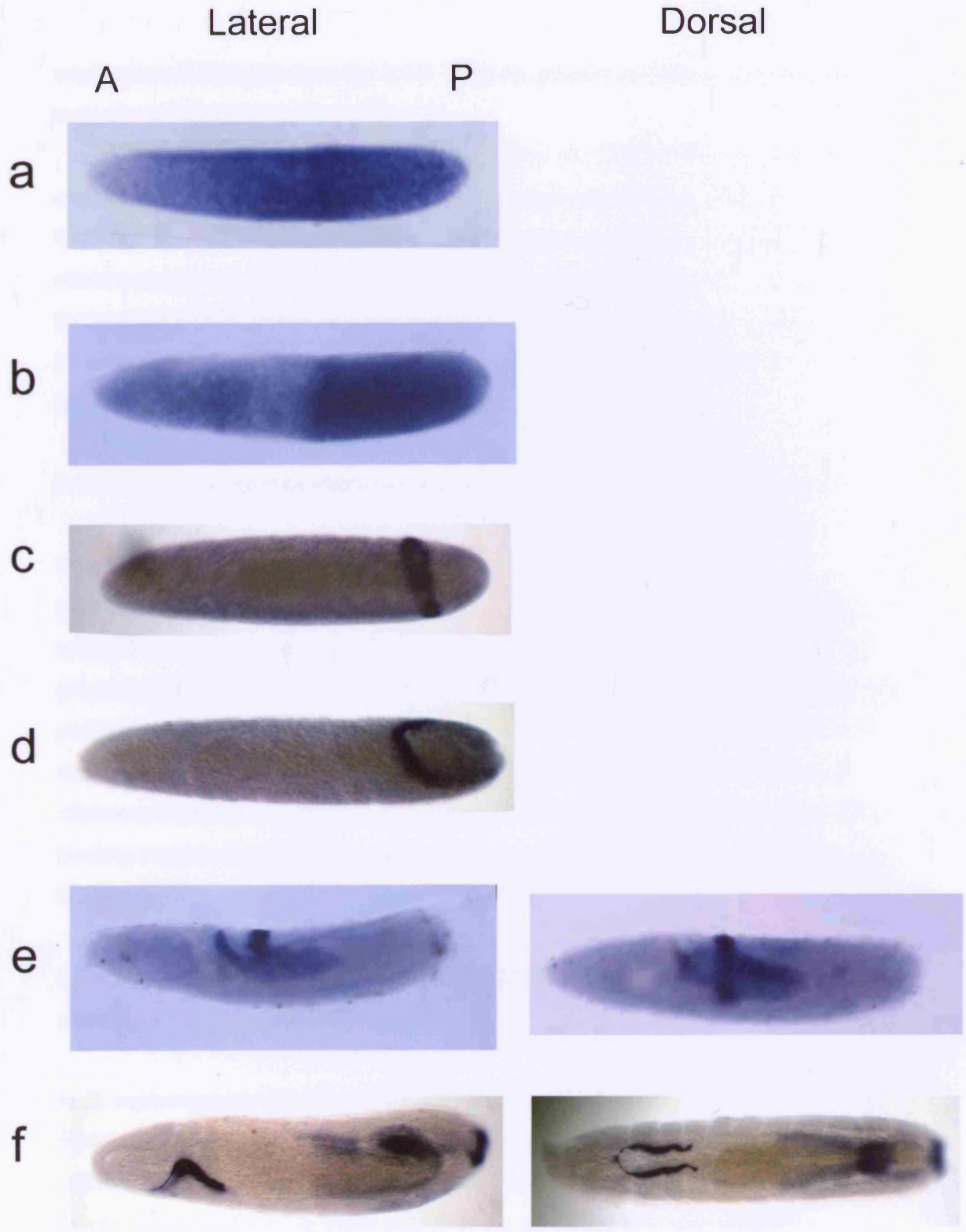


Figure 8.5 A graph showing the simplicity scores of the *cad* coding region and 3'UTR. The sequence position relative to the translation start site is shown on the y-axis. The simplicity score is shown on the x-axis. The structure of the coding region is indicated by the black line; the red box representing the homeodomain. The polyglutamine and polyasparagine runs are indicated by the arrows.



8.7 *cad* mRNA expression patterns in *M. domestica* embryos.
Embryos on the left are viewed laterally, the anterior (A) to the left and posterior (P) to the right and dorsal to the top. Embryos on the right are viewed dorsally, again with anterior to the left. The different stages are as follows:
a. syncytial blastoderm; b. cellular blastoderm; c. gastrulation;
d. germ band extension; e. maximal germ band extension; f. dorsal closure;
g. larval stage.

anal pads will develop (see fig. 8.7f). This expression continues into the first instar larval stage (see fig. 8.7g).

The expression patterns of *M. domestica cad* mRNA are the same as the expression of *cad* in *D. melanogaster* (Macdonald and Struhl, 1986, Mlodzik and Gehring, 1987). This suggests that *cad* regulation of posterior development is conserved between these two species. The expression patterns of *tll* and *hb* in the posterior of *M. domestica* embryos are similarly conserved and both are involved in the regulation of zygotic *cad* expression in *D. melanogaster* (Bonneton *et. al.*, 1997; Sommer and Tautz, 1991; see 3.3.4).

8.2.5 Evidence for translational regulation of *M. domestica cad* mRNA

The *M. domestica* Cad homeodomain is nearly identical to *D. melanogaster cad* (see 8.2.3). Therefore, the presence of Cad in the anterior of *M. domestica* embryos would be likely to disrupt development (see 8.1.2). As *cad* mRNA is present throughout the *M. domestica* embryo during the syncytial blastoderm it is probably translationally repressed in the anterior. Indeed, there is evidence for translational regulation of *cad* mRNA in more ancestral species, including *T. castaneum* (Xu *et. al.*, 1994; Schulz *et. al.*, 1998). The conservation of an RNA binding motif in the Bcd protein between the two species is further evidence for Bcd regulation of *M. domestica cad* mRNA (see fig 1.7; Niessing *et. al.*, 2000).

8.2.6 Comparison of *cad* mRNA secondary structures between *M. domestica* and *D. melanogaster*

In *D. melanogaster* it has been shown that the regulation of *cad* mRNA involves Bcd binding to the Bcd responsive element (BRE) within the 3' UTR of the *cad* mRNA (Dubnau and Struhl, 1996; Rivera-Pomar *et. al.*, 1996). What is the nature of the BRE and is there a BRE in the *M. domestica cad* mRNA?
RNA function is often determined by its secondary structure as RNA binding proteins, such as Bcd, interact with the mRNA via stem loop structures present in

the 5' or 3' UTR (Murata and Wharton, 1995). It is possible to predict the structure of the *D. melanogaster cad* mRNA using the Mfold programme and identify the structure of the BRE (see 2.2.12). The Mfold programme predicts up to 50 structures with the lowest folding energy and therefore the most stable structures. This was done for *D. melanogaster cad* mRNA (See fig. 8.6). In the region of the BRE, a double stem loop structure was present in 23 out of the 25 most stable predicted structures. The majority of the 3' UTR sequence is predicted to fold into the same stable structure in all 25 foldings. In this structure the 5' cap is positioned near the BRE sequence, irrespective of the BRE structure and as a result a Bcd protein bound to the BRE would be positioned close to the 5' cap of the mRNA.

Assuming the *M. domestica* Bcd protein binds the *cad* mRNA it is expected that a similar secondary structure would be found within the *cad* mRNA. Plotting the *M. domestica* RNA secondary structure with Mfold does not produce a double stem loop structure like the one seen in the *D. melanogaster* BRE region. However, the *M. domestica* 3' UTR does form the same stable structure in the 25 foldings of lowest energy. In these structures the 5' cap is positioned close to the start of the 3' UTR as it is in *D. melanogaster* (see fig. 8.6). The position of the BRE close to the 5' cap is important for Bcd repression of *cad* mRNA (see 1.15; Niessing *et. al.*, 2002). Therefore, if a sequence within the proximal 3' UTR of the *M. domestica cad* mRNA is responsive to Bcd regulation its position with respect to the 5' cap has been conserved between the species. The Mfold results indicate that the functional aspects of the secondary structure of the 3' UTR have been conserved despite the increased length of the *M. domestica cad* sequence.

8.2.7 The evolution of the Bcd-*cad* mRNA interaction between *D. melanogaster* and *M. domstica*

The *cad* 3' UTR sequences of *D. melanogaster* and *M. domestica* do not align and there is no recognisable BRE sequence in *M. domestica.* Different RNA

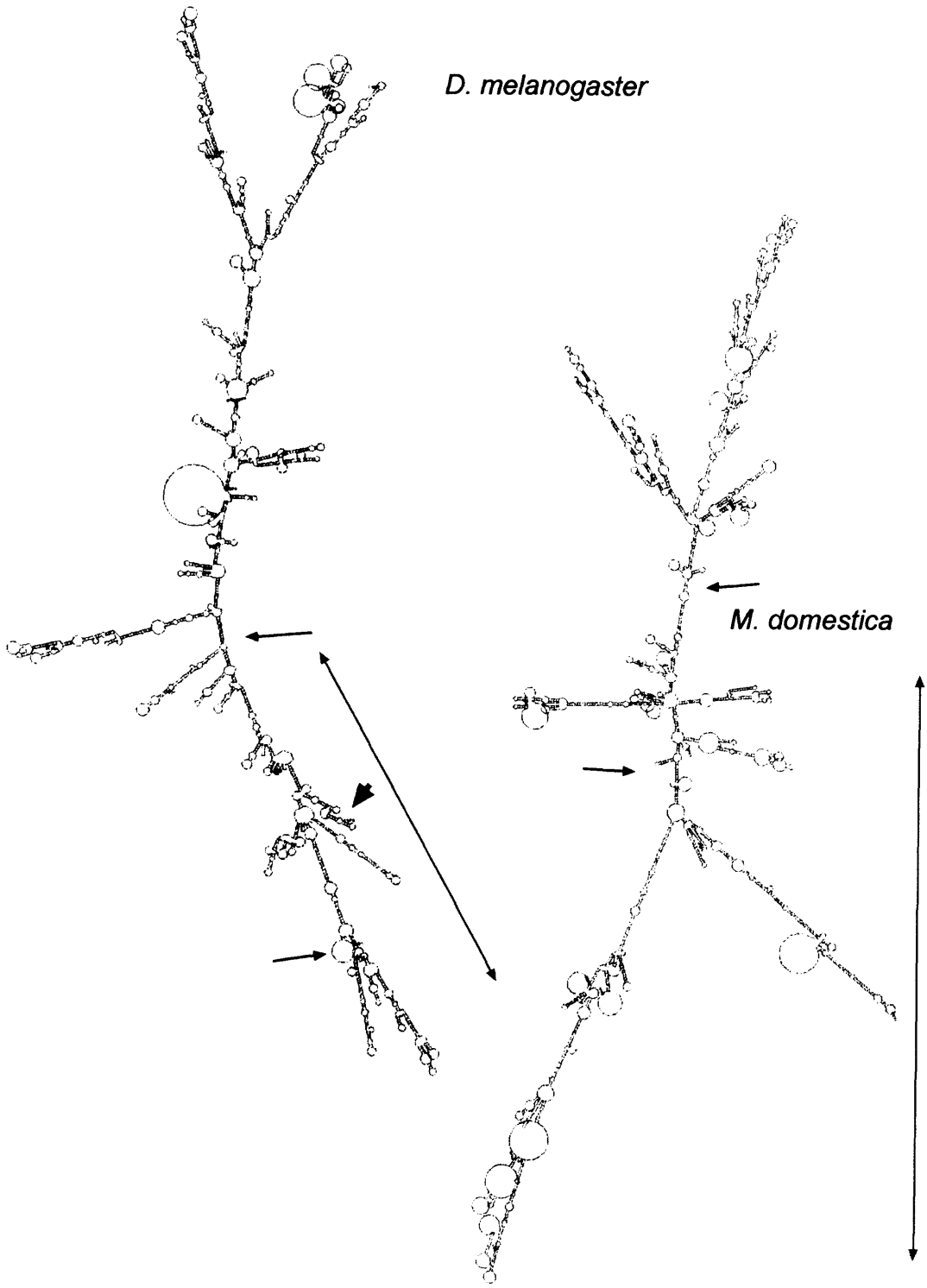


Figure 8.6 Comparison of *D. melanogaster* and *M. domestica cad* mRNA secondary structures. The black arrows indicate the 5' end of each transcript and the blue arrows the start of each 3'UTR. The double-headed green arrows indicate the folding in the 3'UTR that is stable and present in most predicted structures. The BRE is indicated in the *D. melanogaster* transcript by the blue arrowhead.

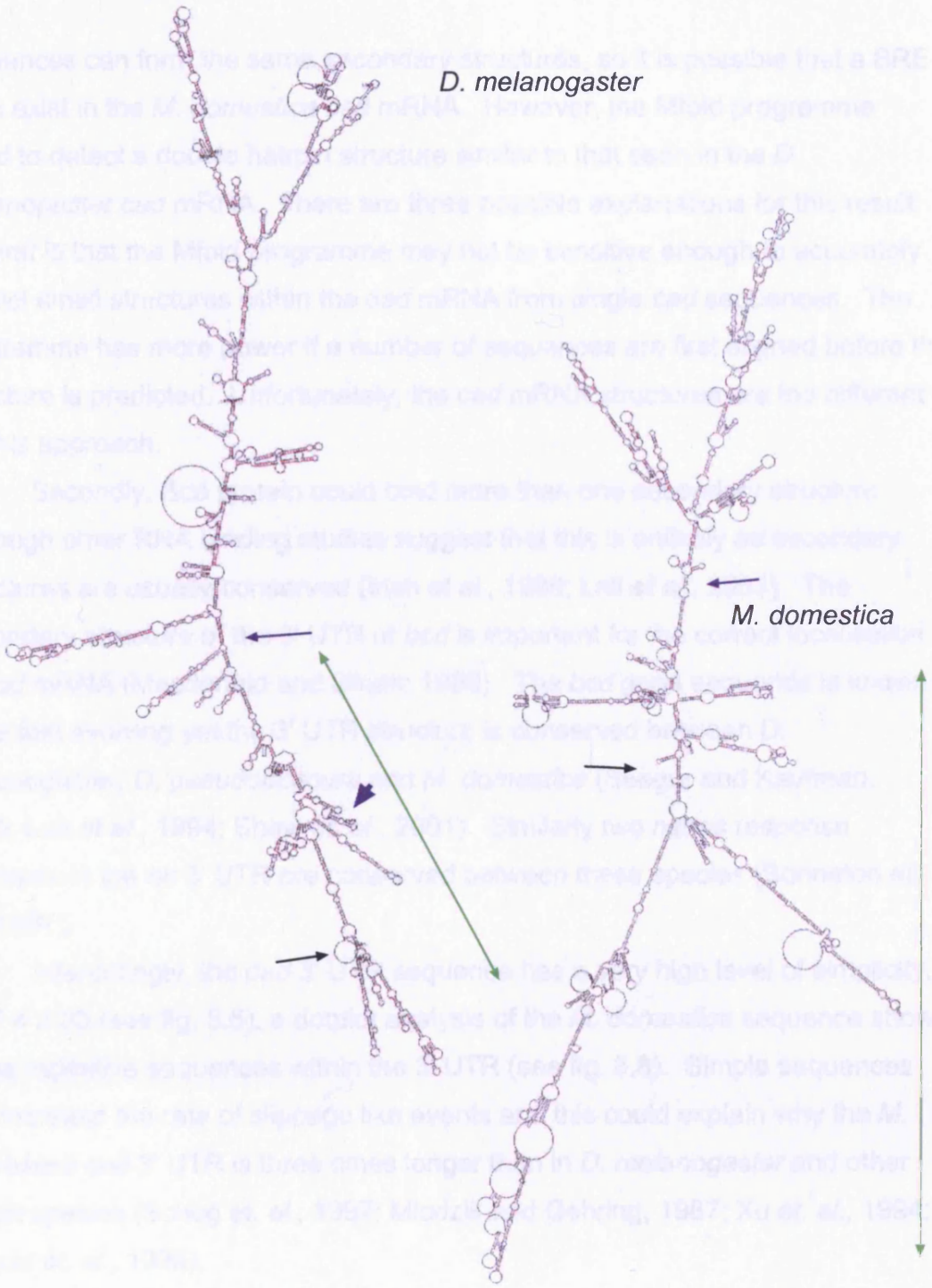


Figure 8.6 Comparison of *D. melanogaster* and *M. domestica cad* mRNA secondary structures. The black arrows indicate the 5' end of each transcript and the blue arrows the start of each 3'UTR. The double-headed green arrows indicate the folding in the 3'UTR that is stable and present in most predicted structures. The BRE is indicated in the *D. melanogaster* transcript by the blue arrowhead.

sequences can form the same secondary structures, so it is possible that a BRE does exist in the *M. domestica cad* mRNA. However, the Mfold programme failed to detect a double hairpin structure similar to that seen in the *D. melanogaster cad* mRNA. There are three possible explanations for this result, the first is that the Mfold programme may not be sensitive enough to accurately predict small structures within the *cad* mRNA from single *cad* sequences. The programme has more power if a number of sequences are first aligned before the structure is predicted. Unfortunately, the *cad* mRNA structures are too different for this approach.

Secondly, Bcd protein could bind more than one secondary structure although other RNA binding studies suggest that this is unlikely as secondary structures are usually conserved (Irish *et al.*, 1989; Lall *et al.*, 2003). The secondary structure of the 3' UTR of *bcd* is important for the correct localisation of *bcd* mRNA (Macdonald and Struhl; 1988). The *bcd* gene sequence is known to be fast evolving yet the 3' UTR structure is conserved between *D. melanogaster, D. pseudoobscura* and *M. domestica* (Seeger and Kaufman, 1990; Luk *et al.*, 1994; Shaw *et. al.*, 2001). Similarly two *nanos* response elements in the *hb* 3' UTR are conserved between these species (Bonneton *et. al.*, 1997).

Interestingly, the *cad* 3' UTR sequence has a very high level of simplicity: RSF = 2.30 (see fig. 8.5), a dotplot analysis of the *M. domestica* sequence shows these repetitive sequences within the 3' UTR (see fig. 8.8). Simple sequences can increase the rate of slippage like events and this could explain why the *M. domestica cad* 3' UTR is three times longer than in *D. melanogaster* and other insect species (Schug *et. al.*, 1997; Mlodzik and Gehring, 1987; Xu *et. al.*, 1994; Schulz *et. al.*, 1998).

The slippage of simple sequences in the *cad* mRNA is likely to have occurred in many small steps. Potentially each insertion of sequence would have been rapidly followed by a second compensatory expansion to maintain a stable secondary structure (Hancock and Dover, 1990; Hancock and Vogler, 2000). Indeed the overall structure of the *M. domestica* 3' UTR predicted by Mfold is

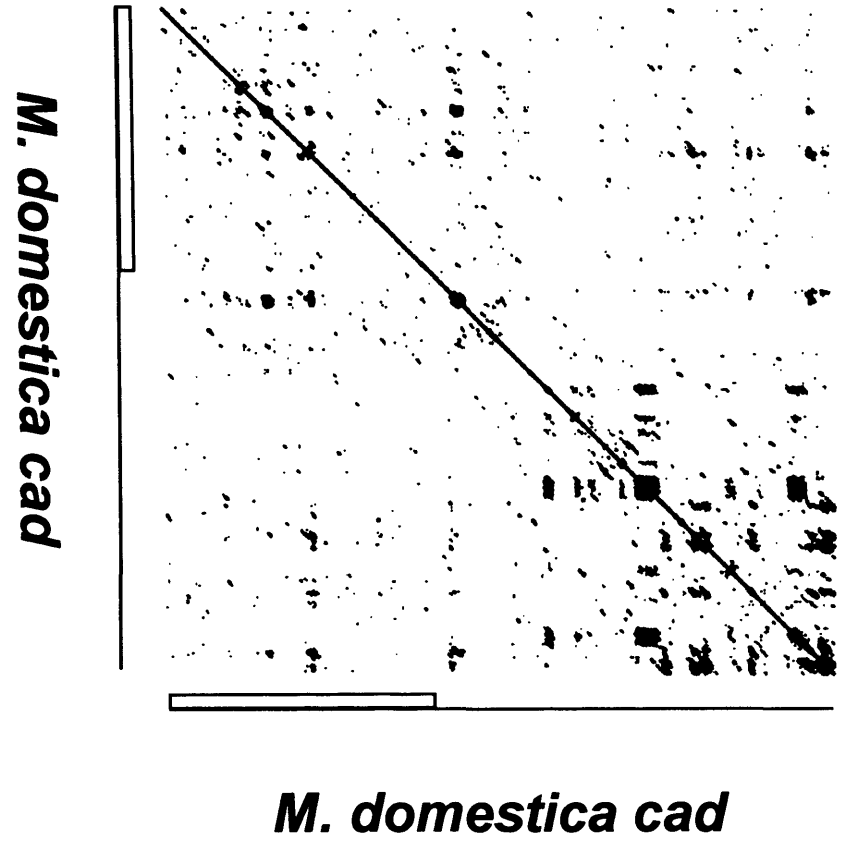


Figure 8.8 Dot-plot of an intra-specific sequence comparison of *M. domestica cad*. This was generated using the dot plot programme (see 2.2.11), a window of 35 bp and a stringency of 18 bases matching. The sequence position is indicated by the transcript structure shown, the black box indicating the coding region.

stable and important functional elements such as the polyA tail are found in a similar position in the structure to those in *D. melanogaster.* If the secondary structure of the *cad* mRNA has changed in shape it is possible that the RNA binding domain of Bcd has adapted to this and this could explain the changes seen between the homeodomains of *D. melanogaster* and *M. domestica* Bcd.

Finally, it is possible that in *M. domestica,* Bcd does not regulate *cad* mRNA. In more ancestral species there is evidence that *cad* mRNA is translationally regulated, however Bcd is absent from these species (Xu *et. al.*, 1994; Shulz *et. al.*, 1998; Wolff *et al.*, 1998; Schroder, 2003). If *M. domestica cad* mRNA is not regulated by Bcd it may be regulated by the ancestral regulatory factor.

In *D. melanogaster* the Bcd protein bound to the 3' UTR of the *cad* mRNA, interacts with the cap associated protein eIF4E to disrupt translation (Niessing *et. al.*, 2002). This motif contains two highly conserved residues yet one is absent in *M. domestica* Bcd (see fig. 8.9; Sachs and Varani, 2000). It is not known whether this change prevents an interaction with eIF4E protein. Interestingly, both the lower dipteran *Megaselia abdita* and *M. domestica* Bcd sequences share the same residue at this position, suggesting this could be the ancestral sequence (see fig. 1.6).

D.melanogaster	YIRPYLP
M. domestica	YIRPYIP
M. abdita	YMRPYIP
eIF4E consensus	Y LX

Figure 8.9 The Bcd eIF4E binding motif and the consensus of eIF4E binding proteins from humans and yeast. X indicates a hydrophobic residue. The non-consensus residue in *M. domestica* Bcd is highlighted in red. *D. melanogaster* – Berleth *et al.*, 1988; *M. domestica* – Bonneton *et al.*, 1997; *M. abdita* – Stauber *et al.*, 1999; eIF4E consensus – Sachs and Varani, 2000.

If the *M. domestica* Bcd protein is unable to interact with eIF4E it could mean that Bcd is not acting as a translational repressor in *M. domestica.* This could explain the absence within the *M domestica cad* mRNA of a secondary structure similar to the one seen in the BRE sequence of *D. melanogaster.*

8.3 Summary

The *M. domestica cad* gene was sequenced and shown to be highly conserved with *D. melanogaster cad* with respect to the protein functional domains and mRNA expression patterns. The *M. domestica* 3' UTR has little resemblance to the *D. melanogaster* 3' UTR. It is apparent that the *M. domestica cad* 3' UTR has expanded in length and this was probably caused by slippage of sequence repeats present within the 3' UTR. Despite the lack of similarity at the sequence level, the 3' UTRs of both species form stable RNA secondary structures in which the start of the 3' UTR sequence is positioned close to the 5' cap. Since the *cad* mRNA expression patterns are the same in both species this suggests the regulation of the *cad* gene is conserved.

Chapter 9 General Discussion: The Evolution of Bicoid Interactions in the Higher Diptera

9.1 Summary of results

The aim of this thesis is to investigate the evolution of the interaction between Bcd and both the *tll* promoter and the *cad* mRNA, between *D. melanogaster* and *M. domestica*. The major findings of the thesis are as follows:

•The *M domestica tll* gene was sequenced and shown to be conserved with the *D. melanogaster tll* gene, with regards to transcript structure and protein coding sequences. The residues unique to the zinc finger of Tll orthologues are conserved in *M. domestica* Tll. The domains of *M. domestica tll* expression are equivalent to *D. melanogaster tll* expression and therefore, the regulation of *tll* is probably conserved between these two species. The onset of *tll* expression is delayed in *M. domestica* in comparison to *D. melanogaster;* this delay is also seen in the expression of *hb* and *otd* and indicates a change in anterior regulation by Tor (Bonneton *et al.*, 1997; McGregor, 2002). The sequence and expression data suggest that the *M. domestica tll* gene is conserved in function with *D. melanogaster tll.*

•The *M. domestica tll* sequences 5' of the coding region were characterised for Bcd binding sites. 33 sequences were protected by the Bcd homeodomain in *DNase*I footprinting experiments in a region extending up to 9 kb from the transcription start site. The Bcd binding site sequences are in accord with the Bcd consensus binding site sequence, but the range of sequences present highlights the flexibility of the Bcd DNA binding domain. Many of the weaker consensus binding sites are found close to strong binding site and the proximity of these sites results in a stronger affinity of Bcd for the weak sites. The *tll* sequences are unalignable between *M. domestica* and *D. melanogaster* and distribution of Bcd binding sites differs completely in terms of number, position and orientation. Comparison of the *D. melanogaster* and *M. domestica tll* promoters suggest there is a difference in the spacing and number of sites between the putative Bcd repressing and activating regions. In the activating

regions there is both a greater number of sites and the sites are more optimally arranged for cooperative interactions (Ma *et al.*, 1996; Burz *et al.*, 1998). It is possible that in the repressing regions the bound Bcd molecules are directly interfering with the transcription machinery (Lee and Young, 2000).

•The inter-strain analysis of *M. domestica tll* sequences shows that there is a higher rate of polymorphism in the non-coding regions than in the coding regions but less than at third position bases of the coding region. This indicates that there are some constraints on sequence change in non-coding regions and indeed the binding sites are 100% conserved. Indels are seen more frequently in the non-coding regions probably because of the limitations on length change in coding sequences. In the promoter region the indels arising between closely spaced binding sites are only 1-2 bp in length and this may indicate the need to keep these sites close together. The same pattern of evolution is seen in the *M. domestica hb* gene sequences. However, the *tll* sequences have fewer indels and base polymorphisms than *hb*, in particular in the coding region. Indel events are partly a result of slippage-like processes that are thought to occur more frequently in regions of high simplicity and indeed the *hb* sequences are more simple than those of *tll* (Hancock *et al.*, 1999; McGregor *et al.*, 2001). The indels in the coding region of *hb* and the 5'UTR of *tll* are present in regions of high simplicity; although, the indels in the promoter regions are found in sequences of both high and low simplicity.

•The results of the band shift assay demonstrate that *D. melanogaster* Bcd has a higher affinity for DNA than *M. domestica* Bcd. The calculation of co-operativity in the binding reactions shows that the *M. domestica tll* promoter is more co-operatively arranged for binding than the *D. melanogaster tll* promoter. Therefore, *M. domestica* Bcd has a greater affinity for the *M. domestica tll* promoter in comparison to the *D. melanogaster tll* promoter than would be predicted by Bcd binding affinity alone. These results agreed with the previous comparisons of Bcd in these two species using the *hb* promoter (Shaw *et al.*,

2002). In general, intra-specific combinations result in a higher affinity interaction than inter-specific combinations. The exception of *D. melanogaster* Bcd and the *M. domestica tll* promoter may be explained by both the high binding affinity of the *D. melanogaster* Bcd protein and the co-operatively arranged *M. domestica tll* promoter.

•A transgenic analysis identified part of the *M. domestica tll* promoter, including regions responsive to regulation by Tor, Bcd and possibly Dl. The expression of the transgene in *D. melanogaster* was generally conserved, including the domains necessary for the correct development of structures missing from *tll* mutant embryos (Pignoni *et al.*, 1990). However, there are differences in expression between the transgene and the wild-type expression of *tll* in *M. domestica.* The most obvious of these is a change in regulation by the terminal system between the two species (Liaw and Lengyel, 1992). This change is likely to be *in trans* because the expression of the transgene appears to resemble that of wild-type *D. melanogaster tll* expression.

•The *M. domestica cad* gene was cloned and shown to have a highly conserved homeodomain in comparison to *D. melanogaster cad*; the rest of the protein was less well conserved and is of unknown function. The *M. domestica cad* mRNA expression patterns are conserved with *D. melanogaster* and more divergent insect species (Mlodzik and Gehring, 1987; Xu *et al.*, 1994; Schulz *et al.*, 1998). The presence of *cad* mRNA in the anterior of blastoderm embryos suggests the need for translational repression. However, a study of the *M. domestica cad* mRNA secondary structure does not provide any positive evidence that this regulation is carried out by Bcd, as is the case in *D. melanogaster.*

9.2.1 The evolution of regulatory sequences

One of the most striking results of the study of the Bcd-*hb* promoter interaction is that, although the promoter sequences are functionally conserved

between *D. melanogaster* and *M. domestica,* they are also unalignable (Bonneton *et al.*, 1997). The same result is found in this study of the *tll* promoter. Such high divergence of functionally conserved regulatory sequences has also been observed amongst other species (for review see Tautz, 2000). In each case the promoters differ in terms of binding site number, arrangement and sequence. Therefore, it is apparent that the promoter as a functional unit can exist in many different forms and importantly can evolve from one form to another whilst maintaining function.

Comparative studies of *cis*-regulatory regions between closely related species give clues as to how such sequence evolution takes place. For example comparisons between *cis*-regulatory sequences of *Drosophila* species show that functional sequences, such as binding sites, are generally conserved whilst the intervening sequences are virtually unalignable (Ludwig *et al.*, 1998; Ludwig, 2002; Kim, 2001; Bergman and Kreitman, 2001; for review see Wray *et al.*, 2003). However, these studies also show that surrounding sequences are not completely free from constraint as the promoter structure and in particular the spacing between binding sites is sometimes conserved (Ludwig *et al.*, 1998; Jenkins *et al.*, 1995). Comparisons between human and rat show a similar pattern of *cis*-regulatory sequence change, as do the inter-specific analyses of the *M. domestica hb* and *tll* genes albeit on a much reduced scale (Dermitzakis and Clark, 2002; McGregor *et al.*, 2001; this work). The divergence of the intervening sequences demonstrate how much mutation and turnover the promoters are experiencing and eventually, over a longer period of time even the conserved blocks of sequence breakdown and the promoters become unalignable (Bonneton *et al.*, 1997; Shaw *et al.*, 2001; Takahashi *et al.*, 1999).

The lack of conservation of the *tll* and *hb* promoter sequences between *D. melanogaster* and *M. domestica* provide evidence of the flexible nature of these sequences and the speed at which changes can arise. Therefore, it is conceivable that the relatively large number of mutation and turnover events in non-coding sequences throw up more novel phenotypes than the same events occurring within coding sequences. The generation of such high numbers of

sequence variants gives an insight as to why evolution of *cis*-regulatory sequences may have played a significant part in the evolution of development between species.

9.2.2 The evolution of binding site sequences

Comparison of the *hb* and *tll* promoters binding site sequences between *D. melanogaster* and *M. domestica* suggest that most of the binding sites have evolved *de novo* since the separation of these species. Is this possible in the 100 MYR since the divergence of *D. melanogaster* and *M. domestica*? It was estimated that a six bp binding site such as a Bcd site can evolve by point mutation in as little as 250 years within a 200 bp region (Stone and Wray, 2001). The calculation did not take into account sequences that were already partial Bcd consensus sequences. Not all sites are functional because sequences flanking the core also influence binding (Shaw *et al.*, 2002). However, this rapid production of binding site sequences provides a constant source of potential Bcd binding sites. Indeed, the generation of a new binding site and breakdown of others has been observed in the *eve* S2 promoter between species of *Drosophila* (Ludwig *et al.*, 1998)

Another source of new binding sites can be from genomic turnover events that can duplicate existing sites. The rapid breakdown of conservation between promoter sequences suggests there is a high rate of turnover in these sequences (McGregor *et al.*, 2001). Mechanisms of turnover such as slippage and unequal crossingover could increase the number of sites present within a local region (Schug *et al.*, 1998). For example, in *C. vicina* there has been a small duplication of a Bcd binding site sequence and both of the resulting binding site sequences are bound by Bcd protein (McGregor, 2002). Evidence from the *Muscoidea* species *Lucilia sericata* also suggests that turnover of sites is occurring. All the higher dipteran species contain Bcd binding sites with the core sequence TAAG, but in the *L. sericata hb* promoter it is the most frequent core sequence present (McGregor *et al.*, 2001). The sequences TAAG and TAATC

interact with the same residues in the Bcd homeodomain, which suggests there is no functional difference between the two sites (see 4.3.3; Dave *et al.*, 2000). Therefore, one explanation for the high frequency of TAAG in *L. sericata* may be such turnover mechanisms and drift.

It is likely that the Bcd binding sites present in both *M. domestica* and *D. melanogaster* have arisen since the separation of the two species by point mutation and turnover of existing sites. Therefore, it is important to understand what affect such sequence changes may have on a promoter output and how these changes can influence the evolution of a regulatory interaction.

9.2.3 Promoter function and the consequences of sequence evolution

A model of promoter function in terms of input (promoter structure) and output (gene expression) was constructed by Gibson (1996) based on the Bcd-*hb* promoter interaction. This model addresses what effect sequence changes such as a gain or loss of a binding site may have on promoter function. The inputs include DNA/protein binding affinities, co-operative interactions and promoter configuration (number of sites). The output or level of gene expression can be given in terms of activation level, activation threshold width (concentration change in which the promoter switches from an 'off' to 'on' state) and position of the threshold (see fig. 9.1; Gibson, 1996). Although this is a simplified model some important points emerge from this study. Firstly, the level of transcription increases with every extra binding site present within the promoter, up to a limit. Interestingly, the *M. domestica* promoters have a greater number of binding sites than the *D. melanogaster* promoters. Secondly, increased affinity to a binding site including co-operative interactions can shift the location of the threshold. Thirdly, the transcriptional response to variations in binding site number, affinity and co-operativity all produce very similar response curves which suggests there are many different permutations of promoter structure which can result in the same output. This final point is illustrated by the Bcd-promoter interactions in *D.*

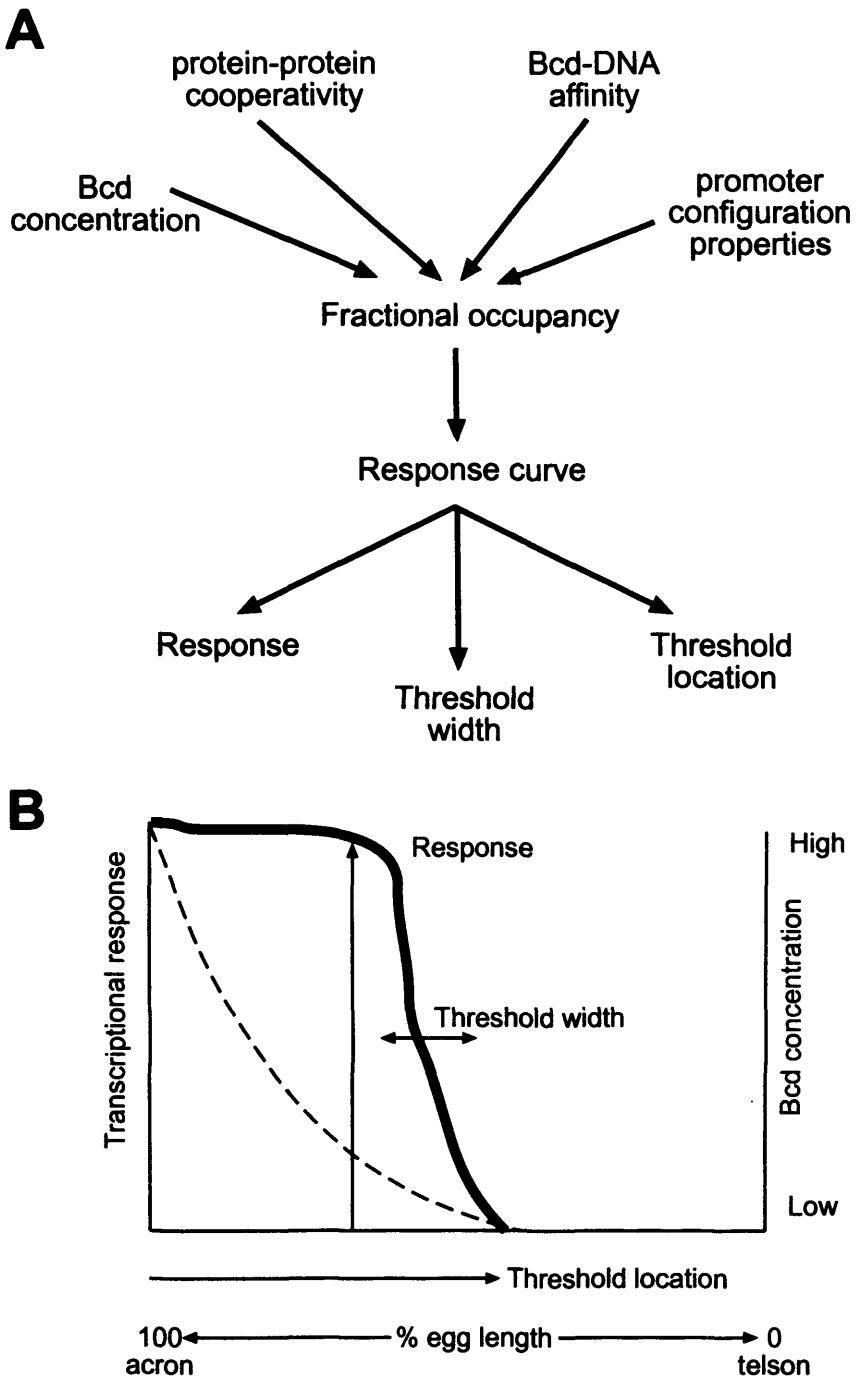


Figure 9.1 Parameters of the Bcd-*hb* interaction (**A**) and response curve (**B**) **A**. The fractional occupancy of the *hb* promoter is determined by the Bcd concentration, co-operativity and binding affinity, as well as the properties of the configuration of binding sites (sequence, spacing, number and orientation). **B**. The fractional occupancyallows the determination of the transcriptional response, threshold width and thresholdposition along the anterior-posterior axis of the embryo. The response curve is representedby the thick line and the Bcd concentration by the dashed line. Adapted from Gibson 1996.

melanogaster and *M. domestica*, which have a different molecular structure but a similar functional output.

Gibson's model could help explain the results of the band shift assays if the results are considered in terms of inputs such as binding properties of Bcd protein and the structure of the promoter (Shaw *et al.*, 2002). The results of the band shift assays demonstrate a level of incompatibility between the species as would be expected if divergence between Bcd and the promoter of the other species is occurring. However, the experiment combining *D. melanogaster* Bcd and the *M. domestica tll* promoter produced a greater interaction affinity than both intra-specific combinations. Evidence from the band shift assays shows that the *D. melanogaster* Bcd has a stronger binding affinity than *M. domestica* Bcd but that the *D. melanogaster tll* and *hb* promoters are less co-operative than their *M. domestica* counterparts. The Gibson model predicts that a stronger binding affinity or greater cooperativity increases activation and agrees with the result seen with *D. melanogaster* Bcd and the *M. domestica tll* promoter. This is because the *D. melanogaster* Bcd and *M. domestica* promoter 'inputs' are both more positive than their counterparts in the other species. Interestingly the model would predict an approximately equivalent output from the interactions of Bcd and the promoter of the same species in *D. melanogaster* and *M. domestica*.

The results of the band shift and yeast assays may be evidence of co-evolutionary change between the components of the interaction between the two species. The functional tests of Bcd and the *tll* and *hb* promoter interactions suggests that the *M. domestica* system may have a more co-operative basis to activation, whereas the *D. melanogaster* system may rely on a greater binding affinity of the Bcd protein. What mechanisms could have resulted in the observed changes in the Bcd interactions between *D. melanogaster* and *M. domestica*?

9.2.4 Evolution of the Bcd protein function and embryo size

The finding that the *M. domestica* Bcd interactions may be more co-operative than those of *D. melanogaster* could be explained by the differences in the physical properties of the embryos of both species. Bcd protein is made at the anterior tip of the embryo and subsequent diffusion to the posterior, combined with degradation of the protein, creates a Bcd concentration gradient in *D. melanogaster* (see 1.9; Driever and Nusslein-Volhard, 1988a). *In situ* hybridisation data and conservation of the PEST domain in *M. domestica* Bcd strongly suggests a similar developmental mechanism is used in *M. domestica* embryos (Sommer and Tautz, 1991; Shaw *et al.*, 2001). However, *M. domestica* embryos are twice the length of *D. melanogaster* embryos, which suggests that the Bcd gradient may be shallower in *M. domestica* due to the trade-off between the diffusion and degradation rates of Bcd protein. This would result in lower levels of Bcd protein throughout the *M. domestica* embryo. As co-operativity is more effective than binding affinity at lower concentrations this could provide an explanation for the selection of potentially greater co-operativity of interactions in *M. domestica* (Gibson, 1996; Burz *et al.*, 1998).

The embryos of the Muscoidea species *C. vicina* and *L. sericata* are also larger than *D. melanogaster*; for example, *C. vicina* embryos are approximately three times longer than those of *D. melanogaster*. Furthermore, in *C. vicina* Bcd mRNA is present in a much smaller anterior domain than in the other species (Schröder and Sander, 1993). Both of these factors could result in a shallower gradient of Bcd protein in *C. vicina* embryos. The homeodomains of *C. vicina* and *L. sericata* Bcd proteins resemble those of the *M. domestica* Bcd sequence in 4 out of the 5 residues that vary between *M. domestica* and *D. melanogaster* (see fig. 1.7; McGregor *et al.*, 2001). This suggests that such changes in the Bcd homeodomain are linked to the difference in co-operativity observed between *D. melanogaster* and *M. domestica*. Indeed, both the *C. vicina* and *L. sericata hb* promoters contain a greater number of binding sites than the *D. melanogaster* promoter and these are suitably arranged for co-operative binding (McGregor *et*

al., 2001). However, it is possible that *D. melanogaster* Bcd is as co-operative as *M. domestica* Bcd but that this property is latent because the promoters in *D. melanogaster* are arranged with less potential for co-operative interactions.

Finally, it should be remembered that other factors are involved in the activation of these promoters, such as the co-factors Chip and dSAP18 and $TAF_{II}110$ (Torigoi *et al.*, 2000; Zhu *et al.*, 2001; Sauer *et al.*, 1996). It is possible that the relationship between any of these factors and the Bcd proteins or Bcd target promoters could result in differences in activation of the *hb* and *tll* promoters between *D. melanogaster* and *M. domestica*.

9.2.5 Evolution of the Bcd network and Bcd function

The functional expression domains of Bcd are conserved between the species *M. domestica, C. vicina L. sericata* and *D. melanogaster*. However, unlike the other two Muscoidea species, *C. vicina* anterior cytoplasm shows no rescue of *D. melanogaster bcd* mutant embryos (Schröder and Sander, 1993). This indicates a difference in the regulation of Bcd targets between *C. vicina* and *D. melanogaster*. Moreover, in *C. vicina* the expression domain of maternal *bcd* mRNA is much reduced and this suggests that Bcd might have a reduced role in anterior patterning. In *T. castaneum* the role of Bcd is performed by Otd in the head region and Hb in the thorax (see fig. 1.4; Schröder, 2003). Perhaps, then, Hb plays a more important role in patterning of *C. vicina* embryos. Certainly in *D. melanogaster,* the over-expression of *hb* can rescue most structures in *bcd* mutant embryos and there is evidence that Hb enhances the action of Bcd in the anterior of wild-type embryos (Wimmer *et al.*, 2000; Simpson-Brose *et al.*, 1994). In *M. domestica* an RNAi experiment indicates a role for *bcd* in head development but a reduced role in development of the thorax in comparison to *D. melanogaster* (Shaw *et al.*, 2001). It is possible that changes in the homeodomain could be linked to changes in the role of *bcd* in development. Apart from the changes in the homeodomain sequences between *D. melanogaster* and the *Muscoidea* there is also a serine

rich domain present in the Bcd proteins of the *Muscoidea* species, which could have an affect on the function of Bcd (Janody *et al.*, 2000; McGregor *et al.*, 2001).

9.2.6 The emergence of Bcd and the effect on sequence evolution

Bcd was chosen as a suitable protein for the study of co-evolution because there were a relatively large number of changes in the homeodomain between *D. melanogaster* and *M. domestica*, which might have been the result of positive selection. In addition, the Bcd-*hb* promoter interaction had been well studied in *D. melanogaster* (see fig. 1.7; Bonneton *et al*, 1997). Unfortunately, at the time it was not known that Bcd was the result of a recent duplication of the *Hox3* gene in the dipteran lineage (Stauber *et al.*, 1999; Brown *et al.*, 2001). After the duplication, the two paralogues *bcd* and *zen* appear to have undergone a subfunctionalisation event (Stauber *et al.*, 2002; Force *et al.*, 1999). An indirect result of this subdivision would be a relaxation of the sequence constraints on both genes with the eventual change in both *bcd* and *zen* sequences (Stauber *et al.*, 1999). There is some evidence that in *D. melanogaster bcd* there has been relaxed purifying selection on functionally unimportant regions of the gene (Baines *et al.*, 2002). Therefore, it is possible that some of the changes seen between the *D. melanogaster* and *M. domestica* Bcd genes are a result of a relaxation of selection.

Our understanding of the evolution of the Bcd network between species of the higher diptera could be helped by knowledge of the function of the ancestral Hox3 protein and the steps involved in the gain of the new role of *bcd* as the anterior determinant. Research into the development of lower dipteran species is underway and so far have shown that the evolution of *bcd* function involved a change in the DNA recognition domain of the protein and gain of RNA binding ability (Stauber *et al.*, 2002). Understanding the gain of *bcd* function may also explain the evolution of the long germ band mode of embryogenesis in which the *bcd* network plays a major part (see 1.14).

9.2.7 Evolution of the Bcd network and RNA binding function

After duplication of the *Hox3* gene, divergence of the *bcd* sequence resulted in a change from a glutamine residue to lysine at position 50 of the homeodomain and this one difference altered the DNA binding preference of Bcd to TAATC. However, comparing *bcd* sequences with *zen* and the *Hox3* genes of lower dipterans, suggests a number of changes would have been necessary to evolve the RNA binding function, including the change to a lysine at position 50 (see fig. 9.2; Stauber *et al.*, 2002; Niessing *et al.*, 1999; 2000). Therefore, the RNA binding ability has almost certainly arisen since the evolution of Bcd as a lys50 homeodomain transcription factor. Perhaps then the additional changes seen in the Bcd homeodomain between *M. domestica* and *D. melanogaster* are related to the evolution of the RNA binding function of Bcd.

```
D.mel Bcd     QVKIWFKNRRRRHKIQS
M.dom Bcd     .................
M.abd Bcd     ............F..EQ
D.mel Zen1    ......Q...MKF.KDI
M.abd Zen     ......Q...MKS.KDR
E.liv Zen     N..V..Q...MKQ.KDM
C.alb Zen     .I....Q...MKENKSN
T.cst Zen     .I....Q...MK..KDQ
Hox3          .I....Q...MKY.KDQ
```

Figure 9.2 Evolution of the Bcd RNA binding *cad* function.
Partial alignment of Bcd and Zen homeodomains in insect species, spaces indicate conserved residues. D. mel – *D. melanogaster* (Berleth *et al.*, 1998), M.dom – *M. domestica* (Shaw *et al.*, 2001), M.abd – *M. abdita* (Stauber *et al.*, 1999), E.liv – *E. livida* and C.alb – *C. albipunctata* (Stauber *et al.*, 2002), T.cst – *T. castaneum* and Hox3 – *Hox3* paralogy group consensus sequence (Falciani *et al.*, 1996).
The residues of the Bcd homeodomain necessary for RNA binding function are indicated along with the equivalent residues of Zen and Zen-like genes of lower dipteran and non-dipteran species. Residue 50 shown in red is necessary for DNA binding function in all proteins and RNA binding function in *D. melanogaster* Bcd. Residue 54 shown in blue is necessary for RNA binding function in Bcd. The *D. melanogaster* residues that resemble the RNA binding motif of HIV Rev-1 protein are shaded in green (Niessing *et al.*, 2000).

Bcd binds to the 3'UTR of the *cad* mRNA and a comparison of the *cad* mRNA structure between *D. melanogaster* and *M. domestica* reveals that there is

little similarity between the *cad* 3'UTR secondary structures of these species. Indeed, the *M. domestica cad* 3'UTR is much longer than that of *D. melanogaster* and the other known insect *cad* genes (Mlodzik and Gehring, 1987; Xu *et al.*, 1994; Schulz *et al.*, 1998). Any change in the overall structure of the *cad* 3'UTR, in particular to the Bcd binding element, could have resulted in a selective change in the RNA binding domain of the Bcd protein. Another possibility is that the lack of evidence for a Bcd responsive element in the *M. domestica cad* mRNA is because this interaction is not conserved between the species and the differences in the Bcd homeodomain are a result of the evolution of this interaction in *Drosophila*. Key to resolving this issue will be the identification of the regulator of *cad* mRNA translation in lower dipteran species and determination of the relationship between *M. domestica* Bcd and the *cad* mRNA.

9.2.8 Understanding the evolution of an interaction in the context of development

It was discovered that the concentration of Bcd could vary up to 30% and still activate *hb* up to the same position along the anterior-posterior axis of the egg (Houchmandzadeh *et al.*, 2002). It has been suggested that this was due to the buffering mechanisms present in development, although what these precise mechanisms are remains unknown (Wilkins, 1997). Such a situation raises the question of whether small changes in the binding affinity of the Bcd protein, in the binding sites or in the degree of co-operativity between sites, would actually produce a selectable difference in activation. It is possible that a number of differences could accumulate in the promoter sequences, both through point mutation and genomic turnover events and be tolerated because of buffering mechanisms (Dover and Flavell, 1984; Small *et al.*, 1992; Ludwig *et al.*, 2000). Such variants might then spread through the population by drift and mechanisms such as gene conversion rather than selection. However, a point might be reached when the accumulated changes in promoter structure and sequence become detrimental to function in which case selection could work against such

changes or promote co-evolutionary changes elsewhere to preserve function (Simpson, 2002; Ruvinsky and Ruvkun, 2003). The complexity of the Bcd network suggests that the former outcome would be the more likely because of the need to keep all interactions properly functional. However in the Bcd interactions studied in this thesis and elsewhere (Bonneton *et al.*, 1997; McGregor *et al.*, 2001; Shaw *et al.*, 2002) the consistent functional changes observed in the components of both the *hb* and *tll* interactions between *D. melanogaster* and *M. domestica* suggest the involvement of positive selection.

The study of Houchmandzadeh and co-workers (2002) raises issues over how little we understand of the mechanisms of development and the extent to which redundancy and buffering play a part (Small *et al.*, 1991; Schaeffer *et al.*, 2000; Wimmer *et al.*, 2000). Before we can properly model the evolution of networks such features of the developmental process must also be taken into consideration (Dover, 2000; Arthur, 2002). Arguably, Bcd interactions although interesting for their complexity are complicated by factors such as the recent origins of *bcd*, the combined functions of the protein and the role of Bcd as a morphogen. Therefore, in continuing the study of the evolution of interactions there is a need to examine a whole range of interactions including those with limited variables to control. For example, choosing and manipulating a transcription factor that activates transcription only above a certain concentration or that is unable to bind in a cooperative manner. In addition the use of visible phenotypes such as the presence or absence of bristles could make co-evolutionary studies more accessible (Sucena and Stern, 2000; Skaer and Simpson, 2000).

9.3 Future work

Differences have been found between the Bcd-promoter interactions of the hb and *tll* genes of D. *melanogaster* and *M. domestica.* One such difference is the apparent divergence in the way Bcd activates target promoters. The functional assays demonstrate that *M. domestica* Bcd shows a weaker activating ability but that the interactions with the target promoters studied are enhanced by co-operative interactions. This leads to a testable prediction that the arrangement of sites within a *M. domestica* promoter should be more sensitive to change than a *D. melanogaster* promoter.

It would be interesting to examine the interaction of Bcd with the *kni*64 promoter, which is known to function co-operatively in *D. melanogaster* (Burz *et al.*, 1998). If *M. domestica* Bcd has been selected to be more co-operative it would be predicted that the *M. domestica kni* promoter would be bound with an even greater degree of co-operativity than *tll* or *hb,* so should show increased sensitivity to changes in the organisation of binding sites. Experimentally, this would involve the sequencing of the *kni* promoter in *M. domestica* and identification of the arrangement of Bcd binding sites. The promoter could then be tested functionally by the deletion/addition or rearrangement of the Bcd binding sites (Ma *et al.*, 1996; Burz *et al.*, 1998).

Another experiment that could detect differences between the Bcd interactions of *M. domestica* and *D. melanogaster* would be to functionally characterise the changes in the Bcd homeodomain between the two species. Potential co-evolutionary changes need to be examined in both interacting components, in this case the homeodomain and its target promoter. To discern which parts of the Bcd protein may cause differences in DNA binding affinity and co-operativity of the protein, chimaeric proteins could be tested in a functional assay (Treisman *et al.*, 1989; Zhao *et al.*, 2000). For example, a chimaeric protein of the *D. melanogaster* Bcd homeodomain and *M. domestica* Bcd flanking sequences could be tested against a reciprocal protein of *M. domestica* homeodomain and *D. melanogaster* flanking sequences. Alternatively, individual

residues could be altered in one Bcd protein to resemble those present in the other species (Janody *et al.*, 2000).

A possible explanation for the enhanced co-operativity of Bcd interactions in *M. domestica* is that the Bcd protein gradient is shallower due to the larger size of the eggs and therefore co-operativity has a greater effect on activation. To test if this is true the Bcd protein gradient should be measured in *M. domestica* (Driever and Nusslein-Volhard, 1988a). To do this it would be necessary to generate an antibody to the *M. domestica* Bcd protein.

It may be the case that the changes in the Bcd homeodomains are related to the interaction of Bcd with the *cad* mRNA. A *M. domestica* Bcd antibody could be used to show that there is an interaction with *cad* mRNA and so identify that part of the mRNA to which Bcd binds (Rivera-Pomar *et al.*, 1996; Dubnau and Struhl, 1996). Confirmation of the Bcd-*cad* mRNA interaction in a non-drosophilid species would also help to determine when the mRNA binding function of Bcd arose.

Appendix 1- *tll* sequence

Length: 11721
Transcription start: 9491
Translation start: 9689
Intron: 9765-9975
Stop codon: 11226

```
     1  CTGCAGGAAT TCGGATCCGC AAGTAAAAAT ATTTGGGTCT TTTGACCCAA

    51  ATATTTTAAC AGTATTCAAT TTTATTCCGA ACATAGGTCA AAACGTTATA

   101  GCAAAGGTTA ACCATTTAAT AGAACAAATA TGACAATACA AAGCTAACAT

   151  TTCAATTGAA TAAAGTGATT CCCAAAAATA ACTTGTATTT ACATATATTT

   201  TAAATTATCA CACATCTTTC AGCATCCTTT GATGGGAACT TTTCAATATT

   251  TCCCACGTTC ACCTGGGCTC GGTCATAGCA ATGGCCAGAT GGTCATGTAA

   301  CATTCACCCA CCATCAATTG TGATGAATGT aCATCAGTCT CCCATCACTG

   351  CAGTGTTATT AGAATAATTC TCATCATCTA TATGAGTACT ATGCGTCGAC

   401  GATGGTGATT ATGATGTGGA TGATGAGTAA TGAAGGATCA CAGAGGACCT

   451  TATCCCACTA CTGAAACATA GCATCACACG CAAGCTTAAT TGTAAAGGTA

   501  ACACAAAAAT TCCcATGCAG TTTTAATATC AACCCATCAA ATGACTAGTC

   551  ATTTGGTCAT TTAAAGtTGC TCCgCATGCT GCCGTTGTTG CTGGTGGCGG

   601  tGGtGGGtTC ACTTTTCGCC CACTCTATAT TTAATTGTTA GCCCACTGAA

   651  GAGCCTTTTT GTTGAACCAC CCaCCCCTAG TTCTAATAAG TGTTATTAAC

   701  GGCCAAACAA TTGAGGGTGA TCCCTTGTTT CcACGCACCC aCCAACTCTA

   751  AATGAAGTTC TAATAAGAGT GTGGGtaCAA TAGtTTCTCA TTCTGATGTT

   801  AGtTGATTTT GATGAGCATA ATAACGATGA tTGCTATTTT GGGAATATCA

   851  TGTCAAGtAT TTTATTGaCC AAGTTCAAAT AACATTTTTG AATTGtGTGC

   901  GCaTGTCTTT TATTTcAgGc TGTGCGAAAC GtTTCTCaTT tcaTGACATT

   951  GtTTGtGTCT ACCAAAAGGG GGAGtTGATT GtAGTGtTTT TTGTCTAGAA
```

1001 tACCATGtAT TGTTGcTCTT TTATGAtCAG AgGAGGGtGA CAACAGAGTT

1051 TCAAATTAAT TtGAATTTTt ATTTACTTTT TTTCAATTCT TCTCACGATa

1101 TTTAAATGAT CATCATGTTT CAAATTTATA ATTGAAAATA GAATCAAAAT

1151 TCGGAACAAT AGTATAAAGT GGAATGCCAC TTTGTGTGTG TGTGTGTTGT

1201 TTTAGTTAAA AAGTTTTCAT TTACAACTGA CACAGTCTTA GTTTGTTCAC

1251 TTGCACAACA AGTTTATGTT TCGtCAAAAT TTTTTGAACG ACAGCAACTT

1301 TCAATAAATT TGTTGCCGTT CGTATACAAT TCCGTGGTCA TTTATTTTGT

1351 GGTAAATATG AGATATATTT ATTCAAATAC ACATCTTATA CATCCaCTTG

1401 GTTTAATTTA ATCTATAACC ATTAACTGtT tCTTTCTGTT ATAAATTGAT

1451 TGAAAAAGGt TTGTATTTGt ATCCaCCACT ATTCTGCCTC AAAaCTTTcT

1501 GGgACGAACA AAATGGaCCA AAGTGGCTAC TATTGAAGGT AGATTTACAT

1551 TTTTAAATCT tCCTTTTTtG tATTTTTTTT TAATTGACAA CTGCGTTGAT

1601 TCCTCATCCT TTTCCTAAGT TGCTTCTTTA GATTAGCGTG TAATAATCGA

1651 TGATTAAACT AGCCaaTTTA GTTGTtATAG AAAGGACCTA GATTtCAACa

1701 CTGTTAGTCA TTTAAAAGTT ATTTCCTCTT GTTTTGtAGT GtCCTGTTAA

1751 ATTAAAGACA GAACCTGCAT TGCAGGGCAA CTCCAAGAGC CATTTGGCAA

1801 ACTTTCTGAA CAATGAATAA CTTTAGAACA TTCGCTGTCT TTAAATACCT

1851 CAATGCGTTA ATTGTCCAAC AAACCTTAAC GTTTCTATTG ATAGTTtCCA

1901 AGACAGATCT ACAGCTTTAA ACTACTGGCC GAGATAAAAA TAaCAAAATA

1951 ACATCAAACc CCATAATTTT GAACTTTTTT AAATTTAATG AATGAAAGCT

2001 AAGAATGTCG GAAATAGAAT CAAATTTTAG AATAAGTTTA GATTTGtACG

2051 AATCGGTCTA CTTTGGCACG AATTATAGCc TTCAAACGGC CAATGAAACc

2101 GTCACACGcT GCACGAATGT TGCTATGCGG TATTTTGGCC CATTCCCCGC

2151 GgAGTGCTTG CTTGAGGTGA TCgACAATTT GGtATTTTTA GTgCCAATCT

2201 TGCTTTCCCA AATACTCCAG ACACAATAGT CCAACGGTTT GGTATCCGgA

2251 GATTTTGGTC GCCATTGTGC GCTGgAAATG AAGCGAGGAa CCTCATTTTG

2301 tAgCCATTCT TGGGTGACGC GTGCTGAGTC CTGTTGgAAT GTCCAAAGTC

2351 TGCGGCCAAA GTGGTTCCGT GCCCAAGGCT TTAATACaCC CTTCAGAATA

2401 TTCTCGCTAT AATATTCGGC ATTTATTCTG ATGCCACGGT TGATTGATAG

2451 GAGCGGAGAT CGACCATCGG CTGTTATGAC GGCCCACACC ATTACCATGG

2501 CCGGCGCTTG AGTTCTGGTG GCCAACCGAA GGTGCACATT TTCATTTTCA

2551 AGTAAACATG ACCATTTTGT TTGTTTACAA ACTGCTGGAC GACAAATGTT

2601 TTCTCATCGG AGAAAACCAA ATTCTGCAGT TCACCACTTT CGGCCAAGCG

2651 AAGCAACTCT TTAGCTCTTT TAAGTCTATT TTCTTTCtGt AGtGGtGTAA

2701 GTTCTTGCAC TCTTtGAAAT TTTAATGGCT TTAACCCGAG TTCATTTTTC

2751 AATATTTGCc GAATGGAATA TTGCGATATG TtCAGCTCAC GAGCCATTTT

2801 TCGGCCACTG CGACACGGGT TTCTTTACTT TTCGAACCAT TTCTGTTGAT

2851 GTTGTTGTTT TCTTCCGTCC ACTTCCTTGA TGTCAGGCTA CGCTACCAGT

2901 ATCAAGAAAC GAGCGATGGA ACAAGATACA AAAGATTTAT TCACACTTTA

2951 ATGCTGGAGT GCTCTAACCA TACCCACTTG TGTTTTCCAG CCAAATATAA

3001 ATAAGTCTCT CTATCACGCT TGAATTTCAT AACGAATAAT TTTTTTCGTA

3051 AAATTCTCAC CAAAATGCTT TTGTaCGCTT GTAAACAATA TAATAAGCTG

3101 cCACTGGAAT AATTTTAACA GCTATTAGAC GAGTGGCTTG GAAATGACAG

3151 CAGTCTGAAG TTGGCTGCAG TTTTTTCATC TCATCCTGTA TATTAAGTCC

3201 GTAATTCCGT TTGTATCCCC TCGAAATATA AGTATTAGAC CTATTAGACC

3251 CCACAAAGTA TATACATATA TATATTCCTC ATCAGCATAA AATTCTATAT

3301 CAATTTAGCG ATGACATTCG GTCACTttAG CTTACGAACG AAGTGAAGTA

3351 AATTGATAAA ATTTTACTCA GATGTGTTAG GTCTACAGGA CTTTTGGTAA

3401 TAAAAATGTT CATGTTTTGG TATAGCTcCC ATTTTTATGA TTTGtGTCAC

3451 ACTTTTCACT GGAAATTTTT CAAATTTGGA ATtTCAAGTT CCAAAAGGAT

3501 ACCACATCCT GTAACAAAAA AtTGCTTATG CAGGtCTACG TcTTCGTATA

3551 GCCCCAAATA ACAGTACCTC cATATATTTG GTATTTTGAT GATTTTAGCA

3601 ACATAGTTCA CAGGAATTGt TTCGTAAGtC CTTAATGTGT TCTTAGGATA

3651 TGCAGATCTA CATCTTCGTA TAGGACCCAA TATAAGAATG AATTACTATA

3701 TTCGATAGTT TTGTGATTTG AGAATTTGAG AAAATTCCTT CCAAATGGTG

3751 GAGGGTATTC AAAGTTCGGT CTGGCCAAAC TTACAGCTTA CTGTAAATTG

3801 GCAGTTAGTG TTTCCAGTAA GTGAGCaACA TAGCAAATAG AGCTGACTGT

3851 ATAGATCCAA GTTGtGGTAT TGtCGtGATA TAACtGTaCC TCCAGATATA

3901 CgATATTTTG ACGATTTTAG CCACACTTTT AAATATAATT ATTTCCTATT

3951 TGGTATTCGG TATTCATAAT GGACACAAGC TTCCATGATC AAAACTGCGG

4001 TTTTCCcACC TGCGATaCCG CCAAACATTC GCTATTTTAA GACcTCTAGC

4051 TACTTTTTTt CAACTGATTA TCTCAAATTT GGTATTTGGC CTATGCAATT

4101 GCTTATACCA AAATGCATGT TTTCGTATAG ACCCCATATA AAGGCATCTC

4151 TCAATATTCG GgtATTTTGt CGGgtATTAT GTCGGAATTT GGAGTAGGAA

4201 CAATTGCTAC TTCTTCCAAT GACAGAATAA TCCTTTATAG CTGTAaCACC

4251 TAATACTCGG AATTTTACAA AATTATTGCT GCATTTTTTA CGGATTTTTT

4301 TTACATTTGG CAAGTTTCTT TTAAATTGTG GAGGAGATTA AAAGTCCGGT

4351 CATTTCGGAC TTAGCTGCTT TTTTTTTCTT GTTTTTCTTG TTGTGGGGAG

4401 AGATATTACT TCCAAATAAA AAAACATGAT TATTTACATG AAAAACAAAT

4451 GAGCAAAGGC TATAGCCGTG CAGGCCGACC ATGTTATCAC CCCTTAGGTC

4501 CAAATCCATT TTGGCAGACT TCTTTTAGAA ATAATGCTAT ATAGCCCTGC

4551 ACAAATGTTA CGTTTGGTTT TTTGGCCGAG GGCACTTGCC CTgTGTTTAC

4601 GCTCTCTGTC GTGCTGATAG GtATTCAGAa CAATTTCACC CGAAAGTTGT

4651 GGTTCATTTT ACCCCCTATA TATAACTATT GTGTCAAAAT CAAAGGCTTT

4701 AAAAATTTTG TCTCTATCAC AAATTTTATC GCGGATCTTT ATCCCTCTAA

4751 ATTACTTGCA ATTCTTGTAT AGAATTCCAT GAATATCTGT TTAATAATTC

4801 TTGAGTAAAT TTGGTACCCA AATATAACGT TTAAAAAATT TTGCCGTTGC

4851 ACGTCAGCtA TTTTCCATAG CTTAATGCAA ATATAAAATA CAATTTCTTA

4901 ATTTAAAGTA ACTATGACAT CAAATTTCTT TTAGCAAATT CTACGATTTT

4951 AAGATTTACT AAATTCCCAG CCATTTTGTC TTTATCTTCC TTGTTCAGTT

5001 TAACTTGCAT TAGCATGCCT TTTGTCATAG GCTGGCGAAT CTAAAGGTTT

5051 ATTCTCTCGC ATTATTATGC TTAACTATGA ATTTAAAGAG AATTCACTGT

5101 AACCAGGGTA TGATTTAAAT TGTAGTACTT AATCGGCGGA TTAAGTTGGC

5151 AGCGATTAAA CATTCGTTAA TGTTAAGCTA TTTACCAAGT CGGCACAAgc

5201 aaGCcACTCA GCTAAGACAA GCCTGCGCTT AGGCATCCAG TCTGCCATGT

5251 TAAGCTACCT TGTCACGACT AAGTATATTC TTGAGTTTTG TGCCAACAGA

5301 ATCTCTCTCT TTGCTGGTTA GCAGAATCCT CCCAGATTCT GGTAAGGATC

5351 AATTATAGCA AGGCCTGGCG TATATCATGC ACAATTCCTT TGGCTCAGGG

5401 CAGGATTCAT TGGAGGtCCA TCATTTTGGG TTAACAGTAG ATCAgAGACT

5451 TAAGGttGGC ATGCACTGGt TTGAGCATGT AACTAAGTTT CTATAGATTT

5501 TTATATCCAA GCTATTACTT TTGACGGATG AAATATCATC ACACCTATTG

5551 CTTATAATAA GGtCtTATGA TTATaGACTA TaGGAGATTa CATATGATAA

5601 AGAtCTTTGG CTTAATTAAA GAATGTaTaT AAaGATAATA TTTAGGCAAt

5651 TtTTTGtTTT tACTTGtTTa CTAAAAtACA GGAAaCAATA ATCCCAaTAG

5701 CGCTTTATGT TGAATTTTGA AGTAAAAACA TTTTGAAACC TAAAGAATGt

5751 TAGGACACCT TCGATATATA CGATTTCCcT CTGGACTTCT GTGCGAAGAT

5801 GTTTTGATCT TATATAATAA GATATTTTGA TCTTACATGA GAAgAACATT

5851 TTGGtCTTCC ATGAAAAGAC TATTTTGGTC TTCTATGAGA AGACCTTTTT

5901 GGTCTTCTAT GAGAATGCCT TTTGGTTTTC TATGGGAGGA TCTTTTTTGG

5951 CCTTCTATGA AGAAACCTTT TTGTTCCATC ATGGTCTTCC ATGAGAACAC

6001 CTTCGGGCCT TCTATTGAAA AATTTCTGAT CTTGTATGGA AGAACTTCTG

6051 ATCTTCTATG AAAGACCTTC TGATCTTCAA TAGAAGACCT TCTGGTCCTC

6101 TATGGAAGAC CTTCTGGTCC TCTATGAGAA AACTTTTTTG GCTTCTATGA

6151 GAAGACCTTT TTGATCTTCT ACGAGAAGAC ATTCGGGCCT TCTATGAGGA

6201 GACCTTCGAG CCTTCTATGA GAAGACCTTC AGGCCTTCTT TTAGAATGTC

6251 ATCTTCTATA GGAAGATCTT CTATAAGAAG ACCTTCGTGT CATCTAAAAA

6301 ATATTTTCAT CTTCTGAAAG AAGAATGCTT TCCTGCCGCC CATAAAATGA

6351 CTTTCTTGTC TACTATGAGA ATACCTTCTG ATTTTCTATG AAAAtAACTT

6401 TTGGCCTTCT TCAAAAATAC CTGGTCATTT ATAATAAGAT GGACTAGTAG

6451 ATCTTCTTTA GTAGTAAAAG TAGTCCTACA TGAGTAGCAA AAATACTGtT

6501 AAAAATCTCC ATGGATACAT GAATAACAGT CGAAATTTGA TTTTTATACC

6551 CTTCACCATT GTAACACCTC GAAATATATA TTGTAGACCC CACAAATTAT

6601 ATATAATCTT GATCAGTATA AAATTCAATG TCGATTTAGC AGTTTCCGTC

6651 CGTCTATCCA TCTGTGGAAA TCACTCTAGC TTCGAAACGA ATTGAAGTAG

6701 ATTGATAAAA TTGTTCTCAA ATGCAGGAAT TTTGGTAGTG AAAATGGGAC

6751 ATGTAGGTCC ACGTTTTGGT ATAGCCCCCA TATAACGGTA CCTCCCGATA

6801 TTCGGTATTT TGATGATTTT AGCGACATTA TTCACCGGAA TTGTTTCAAA

6851 TTTGGTATTA GAAGTTCGAA ATGGATACTA CAGCCTAAGA CAGAAAAGTA

6901 CCTCGCGATA TTCGGTACTT TGTTCATTTT AGCAACATTA TTCACCGGTA

6951 TtATTTCACA TTTGGTATTT AAtGtTCGTA AtGGATGCAC AGGAGATTGT

7001 CTGTGCAGGT TCATGTcTTC GTATAgCCCA CATATAAAGT TACCcTCCCG

7051 AtATTCGGTA TTTTGATGAT TTTAGCGACA TTATTCGCCG ATTTTTTTTT

7101 AAGTTTCGCA TTTGAATATT GTAGTTTCTT ATGCCTAAAG TTACCTTTGC

7151 AGGTCCATTT CGTCGTATAG GCCCTAATAT AAGAGCAACA TTCGGTATTT

7201 TTAAGATTTT AACAACATTT TTAATGGATT TTATTGTTGT TGAGAATTGC

7251 AAAATTCGTT CCAAATGGTG GAGGGTATTC GAAGTTCGGC CCGGCCGAAC

7301 TtACAGCTTT TCTTTTCtTG TTTAAtATAT TTTTAGTAGG CCTTGACAAA

7351 TAAAAGTGAT AACTaAATCA TAGATATCAT TAAAGTtGGC ATAAGAACTT

7401 GAGACTGTTT CGTATAAaCC CCATAtAACG GTTTGCACAG ATCTCAATTG

7451 CGTTAGGTGT ATATTACAAC CAAAAGCGGT GTTGCAGATT CTTAGTGGAT

7501 TGAAATCCGC AATTTGAAAT ACACATGTCT ATATTTGTGG TAGCCTCAaA

7551 TAGAAATTcc AAAAAAAAaa TAaTTtTTAt tAAAgAAAAA TTATAAATTT

7601 TGCTCCACGC TTTGTTTTTT AAAGCCCTAG CTCCTTTTtC CaCGATAGTt

7651 GTTATTTTAc ATTTTAAaAT TGTAATGGGA AGAAAgCTTC TCTGAACAAT

7701 GAAGAAACAT TTAGATGTTT ATAAGTGCAG TTCTAAGATT TAGAAAAAAa

7751 TGCCACAGAA ATGTCCgTTT AtGTTGTCTC ATCTtCTTGT AAGCATATGT

7801 TATTTTTAAA TTACTCAtCG GAATATATCT CAAGTATGAG ATTTGTCAAA

7851 tATTGGTGcC CTTTCTcAAA CCcTGTCTAA ccccctgAtt GTttttATaa

7901 aagtatCAAA TATGGGATAA AGCCTGATTA TTTTAcAAcg caaTataact

7951 acaatattaa aataaataaa tGCaCAaCAC TGTTTTAAAA ACCATTTTAA

8001 TTAAATTCTA AGCTAAgAtC TTTggacTat tCATTAGGAT TTTCAAATCC

8051 AATGgTATTT TTATGACATC GGTCTtCCTA AGTATttttt tTtTTTtTCG

8101 AAATAATCTC TCTTTTGCAT ATTTTGTAAG AATTCCAAAA TAGAACTAAA

8151 CAAATTTGTC TAAATATTGA TATTATGTTT TTATTTTCAG TTGTTATATC

8201 TGTGGTTGAA TGGTTTTGGC ACGCTTGTTC CTAACAATGC TTGTCATTGA

8251 CTTTTAAgAC CCAAGGATCT TCTATTCGGC tAATATTAAT CTTCAAAATT

8301 AAATTCAAGA TAAGATTCAC AAAGATTTTG TTAATTTtAC ATgCGGACTC

8351 ATACAATTTT TTTTAATTTA TTTTATGCCA CTCTGCCGTG TTTTTGtCAA

8401 AAAGTTtGtA AAAACTCGAA ATGATGGtCC TAGATCCaCC ATATTGTGTC

8451 CAATTAATAT ATCTCGaCGa CCAGTCCATT TTAAATTGTA AATACTTCAC

8501 AGGTCGCAGA ATACAATCTT ATGATATTTG CAGATGATAG TGAATTTAGA

8551 AaCTGAGTTC TTAATGTTGA AATATTTCTC ATCAAAATTC ACTATTATTA

8601 TCATTTTTTT CTGAAATATA ATTtGtTGTC ATATATTGCT TCTCATTGCA

8651 AAACCTATTT ACCAATCATA TATGATTTTT TTCTCAACTG AAACGAGGTT

8701 TTTCATAAAA AAACCAAAAT CTTTGCAAAG TTCAAtTGCT TTCCCTTTAA

8751 TCCCAATAAT TACATCAACA CGCTTTGCCC CCAGTACAAA ATGAGAAACA

8801 TTAAGAATTC CAATGTTTGA TTTTTTCAAT AGTTCAGGGA TTATCCaAAA

8851 TAAAGAATAA AATTTAGAAA TTCCCAACTG CATATCAAAC AGCCATCGaC

8901 AATTAATTCC AaGTAAACGA AAAAAAaCTA TATTCCTCTT GaCAATGAAT

8951 GCaAAAAGAG AGCAATTaAG TGAacCAAAG AAAGAAAACA TCATTaGAGA

9001 AAaTTAAAAC GTTAGTAGAA CATCTAGAGC aCCGGAGAAC AGGTaACAAC

9051 AACGATGTaC ATTAACAATG GtCCTTCAGT TCAGTTGATG GAAGAgCAGC

9101 GCACACAGCA CAGTGGCAAT AAACGTGCCA CCCTTTTCTC ACAGGCACCG

9151 GGTACAATAA TTCTTTCATA AAATTTTCGA AATGAGAGCA GAACACAGGT

9201 AAACCACCAT AAAAGAGATG TCATAAAATC ATCacCGGCA CTGAATGAAA

9251 ACCTCTGCCA ATGTATGCGA ATATACACAC GCACAGGCCA AAAGATCTGC

9301 ACAAGGGGTT GTTCCACACA CACGTACACG AATGCTCTTC ATTAAAGGAT

9351 AATGTGAGAG AAAGAGAACA AAAATGCACC CTCTTGGCTG TGTGTGTAAC

9401 CAATGAACGA ACGCTTACTA GGTAAAAGGG GATGATCGTA ACATTGGTTG

9451 GCCGTATGAA TGGTAGAGTG CAAAATTTGC GCCAGGACTA TTAAAGGACG

9501 AGGACGGTGG TGAATTAGCA ACACAAACTA TTTGGATCTC AAACAGTGAA

9551 CACAACTCAA ACAGAGCTGA GAACACTAAA AATtAACAAA ATATCTTTAC

9601 AACAATTaCg AATTAAAAAA TaTTTGATAT AaCaAAAAAC ATaCATTTAa

9651 CCaCGATCAA GGATTaCTTT aCAATAaCAA aCAaCAAAAT GCaAACCACC

9701 GAaGGaTCTC CcGATATTAT GGATCAAAAA TACAACtcCG tCAGATTATC

9751 TcCAGCTGcG TCAAGTAAGT ATTTcAaCCA AGCAATATTc TcAACGAAGC

9801 AaGCGAATTA aCAGCAAGGa TGAaCCAAAA GCTaCTTTGC CATTTCAAAG

9851 GATTACCCAA ACTTGAAAAA GCCAAAGAAG TTTTGcTTCC TCTCCAGCTA

9901 GGAATTCACA AAATATTTCT CGAgTCTCAT TTCACAATTT ACTAATTGAT

9951 TTTTTCTCtC TTGCTCTAAT TtTAGGTCGC ATcCTATATC ATGTgCCTTG

10001 tAAAGTTTGC CGTGAcCAtA GTTCTGGCAA ACATTACGGT ATCTATGCAT

10051 GTGATGGCTG TGCTGGTTTC TTtAAGCGTT CCATTCGGCG TTCCCGCCAA

10101 TATGTGTGCA AATCCCAGAA ACAAGGACTC TGTGTGGTGG ACAAAACCCA

10151 TCGCAATCAG TGCCGTGCCT GCCGCTTGCG CAAATGTTTC GAAGTTGGCA

10201 TGAACAAAGA TGCTGTCCAA CAcGAaCGTG gACCCCGCAA CTCCACATTG

10251 CGCCGCCACA TGgCAATGTA CAAAGATGCC ATGATGGGtG GCTCTgAAAT

10301 GCCCCAGATC CCAGCTGAAA TTCTCATGAA CACcGCAGCT TTAACTGGTT

10351 TCCTCGGCTT GCCAATgCCA ATTCCAGGAT CCCATCACAT GCATCCCAGT

10401 TTGGCTGgAG CCTTCCcTGC ACCACCATCA GTTTTGGATT TATCTGTACC

10451 TCGTGTACCC CAACATCCCA TGCATCAAGC TCATCCCGGT TTTTTTGCAC

10501 CCACTGCCGC cTACATGAAT GCCTTGGCTG CCACCCGTGT CTTGCCACCC

10551 ACACCTcCAc TAATGGCTGC CGAACACATT AAAGAAACTG CAGCCGAACA

10601 TCTCTTCAAG AaCATCAACT GgATTAAGAA CGTACCCTCA TTTGGTGAAT

10651 TGCCACTGcC CGATCAATTG CAGTTGtTGG AAGACTcCTG GAAGGAATTC

10701 TTCATCTTGG CTATGGCGCA ATACCTCATG CCCATGAACT TCACTCAGCT

10751 ATTGTTCGTT TACGAATCGG AAAATCCCAA CCGCGATGTT ACCGGTTTGG

10801 TAACCCGTGA AGTTCATGCC TTCCAAGATG TCCTAAATCA ATTGTGTCAT

10851 CTCAACATTG ATAGCCATGA ATATGAGCTC ATTCGTGCCT TGACCCTATT

10901 CCGCCGCCCT GGCTCCGATG ATTTGGCCAA TTCTTCACTG TCCACCAGCA

10951 ACGGCAGTCC CAACTCCAGC ATCTCTGCCG AATCTCGTGG CCTCATCGAA

11001 AGCACCAAAA TTGCCGCCTT ACATGATGAG AGCCGTAATG CCCTCATTGG

11051 CTACATTGCC CGTcTACATC CCGGCCAACC CATGCGTTTC CAAAGCATTA

11101 TGAGTGTATT GACCCAGATG CACAAAGTCT CCTCGTTTGC CATTGAGGAA

11151 TTGTTCTTcC GCAAAACTAT TGGTGACATT ACCATTGTTC GTcTGATTGG

11201 TGACATGTaC AGCCAGCGAA AAATTTAAAA TGCAACGAAC AGTTTGCCGT

11251 CGGACAAATG ACTTGTGGGC CCAAAGAAAG TTTGGGTCGA TACACCAGTG

11301 AAACATGCaC ACATTGGGGC ACAGGACACT TGAAGTGTCC aCAaCCGCTG

11351 cTAAGCCCAg TGGGGCCCTG ATGCACTCGA TTGCCTCAGG TCAGCGGAGT

11401 GCTCCAGTTc TGGAAACTGT GATAATAATG CCTCAGcCTG aTTTATCATC

11451 GGATATTTGT AGACCAAAAT GGaCAATACT CAAAGTTTTG ATGTGGCCAa

11501 CAGcTGTAAG CACTtAAAAC AAAAaGGCTT CCAaTTtAAA ATAAAAagCC

11551 TTAATGAATG AAATAAgCCT TcAAaGAAAa GTGTGAaCAA AAATATCATT

11601 CCAAGTGATC TTAATCGTAG CTTAAGtTTA AgCAAAAaGt TGCAAAATTT

11651 TTtGtTATAT GgTTTAAAAA GGAGAGgAGG AAAGGAACCA ACTTACCAAA

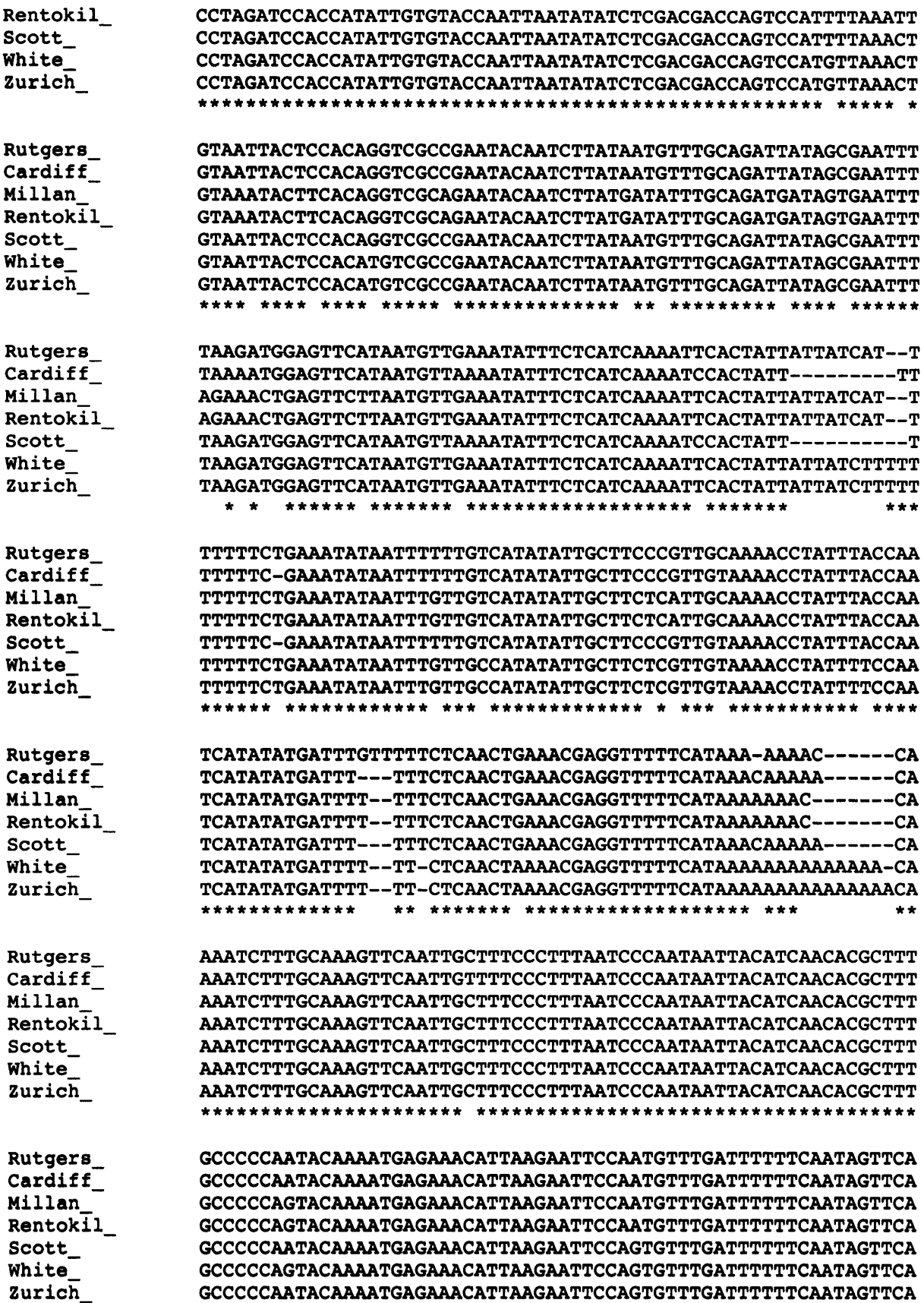
11701 AATGGTGgAA aTCAAGTTAT T

Appendix 2 - *M. domestica* strain comparison

tll promoter alignment (ClustalW)

```
Rutgers_        AGATGTTTATAAGTGCAGTTCTAAGATTTAGAAAAAAATGCCACAGAAATGTCCCTTTAC
Cardiff_        AGATGTTTATAAGTGCAGTTCTAAGATTTATAAAAAAATGCCACAGAAATGTCCCTTTAC
Millan_         AGATGTTTATAAGTGCAGTTCTAAGATTTAGAAAAAAATGCCACAGAAATGTCCCTTTAC
Rentokil_       AGATGTTTATAAGTGCAGTTCTAAGATTTAGAAAAAAATGCCACAGAAATGTCCCTTTAC
Scott_          AGATGTTTATAAGTGCAGTTCTAAGATTTAGAAAAAAATGCCACAGAAATGTCCCTTTAC
White_          AGATGTTTATAAGTGCAGTTCTAAGATTTAGAAAAAAATGCCACAGAAATGTCCCTTTAC
Zurich_         AGATGTTTATAAGTGCAGTTCTAAGATTTAGAAAAAAATGCCACAGAAATGTCCCTTTAC
                ******************************** ********************************

Rutgers_        GTTGTCTCATCTGCTTGTAAGCATATGTTATTTTTAAATTACTCAACGGAAAATATCTCA
Cardiff_        GTTGTCTCATCTGCTTGTAAGCATATGTTATTTTTAAATTACTCAACGGAAAATATCTCA
Millan_         GTTGTCTCATCTGCTTGTAAGCATATGTTATTTTTAAATTACTCAACGGAAAATATCTCA
Rentokil_       GTTGTCTCATCTGCTTGTAAGCATATGTTATTTTTAAATTACTCAACGGAAAATATCTCA
Scott_          GTTGTCTCATCTGCTTGTAAGCATATGTTATTTTTAAATTACTCAACGGAAAATATCTCA
White_          GTTGTCTCATCTGCTTGTAAGCATATGTTATTTTTAAATTACTCAACGGAAAATATCTCA
Zurich_         GTTGTCTCATCTGCTTGTAAGCATATGTTATTTTTAAATTACTCAACGGAAAATATCTCA
                ************************************************************

Rutgers_        AGTATGAGATTTGTCAAAAATTGGTGACCTTTCTCAAACCCTGTCTAACCCCCTGATTGT
Cardiff_        AGTATGAGATTTGTCAAAAATTGGTGACCTTTCTCAAACCCTGTCTAACCCCCTGATTGT
Millan_         AGTATGAGATTTGTCAAAAATTGGTGACCTTTCTCAAACCCTGTCTAACCCCCTGATTGT
Rentokil_       AGTATGAGATTTGTCAAAAATTGGTGACCTTTCTCAAACCCTGTCTAACCCCCTGATTGT
Scott_          AGTATGAGATTTGTCAAAAATTGGTGACCTTTCTCAAACCCTGTCTAACCCCCTGATTGT
White_          AGTATGAAATTTGTCAAAAATTGGTGACCTTTCTCAAACCCTGTCTAACCCCCTGATTGT
Zurich_         AGTATGAAATTTGTCAAAAATTGGTGACCTTTCTCAAACCCTGTCTAACCCCCTGATTGT
                ******* ************************************************************

Rutgers_        TTTTATAAAAGTATCAAATATGGGATAAAGCCTGATTATTTTA--CAACGCAATATAACT
Cardiff_        TTTTATAAAAGTATCAAATATGGGATAAAGCCTGATTATTTTA--CAACGCAATATAACT
Millan_         TTTTATAAAAGTATCAAATATGGGATAAAGCCTGATTATTTTA--CAACGCAATATAACT
Rentokil_       TTTTATAAAAGTATCAAATATGGGATAAAGCCTGATTATTTTA--CAACGCAATATAACT
Scott_          TTTTATAAAAGTATCAAATATGGGATAAAGCCTGATTATTTTATTCAACGCAATATAACT
White_          TTTTATAAAAGTATCAAATATGGGATAAAGCCTGATTATTTTA--CAACGCAATATAACT
Zurich_         TTTTATAAAAGTATCAAATATGGGATAAAGCCTGATTATTTTA--CAACGCAATATAACT
                ******************************************  ***************

Rutgers_        ACAATATTAAAATAAATAAATGCACACACTGTTTTAAAAA-CCATTTTAATTAAATTCTA
Cardiff_        ACAATATTAAAATAAATAAATGCACACACTGTTTTAAAAA-CCATTTTAATTAAATTCTA
Millan_         ACAATATTAAAATAAATAAATGCACACACTGTTTTAAAAA-CCATTTTAATTAAATTCTA
Rentokil_       ACAATATTAAAATAAATAAATGCACACACTGTTTTAAAAA-CCATTTTAATTAAATTCTA
Scott_          ACAATATTAAAATAAATAAATGCACACACTGTTTTAAAAAACCATTTTAATTAAATTCTA
White_          ACAATATTAAAATAAATAAATGCACACACTGTTTTAAAAA-CCATTTTAATTAAATTCTA
Zurich_         ACAATATTAAAATAAATAAATGCACACACTGTTTTAAAAA-CCATTTTAATTAAATTCTA
                ****************************************** ******************

Rutgers_        AGCTAAAATCTTTGGACTATTCATTAGGATTTTCAAATCCAATGTTATTTTTATGACATC
Cardiff_        AGCTAAAAT-TTTGACTA-TTCATTAGGATTTTCAAATCCAATGGTATTTTTATGACATC
Millan_         AGCTAAAATCTTTGGACTATTCATTAGGATTTTCAAATCCAATGGTATTTTTATGACATC
Rentokil_       AGCTAAAATCTTTGGACTATTCATTAGGATTTTCAAATCCAATGGTATTTTTATGACATC
Scott_          AGCTAAAATCTTTGGGAT-TTCATTAGGATTTTCAAATCCAATGGTATTTTTATGACATC
White_          AGCTAAAATCTTTGGACTATTCATTAGGATTTTCAAATCCAATGGTATTTTTATGACATC
Zurich_         AGCTAAAATCTTTGGACTATTCATTAGGATTTTCAAATCCAATGGTATTTTTATGACATC
                ********* ****     ************************** ***************
```

```
Rutgers_      GGTCTCCCTAAGTATTTTTTTTTTTTTTTCGAAATAATCTCTCTTTTGCATATTTTGTAAG
Cardiff_      GGTCTCCCTAAGTATTTTTTTTTTTTTTTCGAAATAATCTCTCTTTTGCATATTTTGTAAG
Millan_       GGTCTCCCTAAGTATTATTTTC-------GAAATAATCTCTCTTTTGCATATTTTGTAAG
Rentokil_     GGTCTCCCTAAGTATTATTTTC-------GAAATAATCTCTCTTTTGCATATTTTGTAAG
Scott_        GGTCTCCCTAAGTATTATTTCT------CGAAATAATCTTTCTTTTGCATATTTTGTAAG
White_        GGTCTCCCTAAGTATTATTTTC-------GAAATAATCTCTCTTTTGCATATTTTGTAAG
Zurich_       GGTCTCCCTAAGTATTATTTTC-------GAAATAATCTCTCTTTTGCATATTTTGTAAG
              ***************** ***         ********** ********************

Rutgers_      AATTCCAAAATAGAACTAAACAAATTTGTCTAAATATTGATATTTTGTTTTTATTTTCAG
Cardiff_      AATTCCAAAATAGAACTAAACAAATTTGTCTAAATATTGATATTATGTTTTTATTTTCAG
Millan_       AATTCCAAAATAGAACTAAACAAATTTGTCTAAATATTGATATTATGTTTTTATTTTCAG
Rentokil_     AATTCCAAAATAGAACTAAACAAATTTGTCTAAATATTGATATTATGTTTTTATTTTCAG
Scott_        AATTCCAAAATAGAACTAAACAAATTTGTCTAAATATTGATATTTTGTTTTTATTTTCAG
White_        AATTCCAAAATAGAACTAAACAAATTTGTCTAAATATTGATATTATGTTTTTATTTTCAG
Zurich_       AATTCCAAAATAGAACTAAACAAATTTGTCTAAATATTGATATTATGTTTTTATTTTCAG
              ********************************************** ***************

Rutgers_      TTGTTATATCTGTGGTTGAATTGTTTTGGCACGCTTGTTCCTAACAGTGCTTGTCATTGA
Cardiff_      TTGTTATATCTGTGGTTGAATTGTTTTGGCACGCTTGTTCCTAACAGTGCTTGTCATTGA
Millan_       TTGTTATATCTGTGGTTGAATGGTTTTGGCACGCTTGTTCCTAACAATGCTTGTCATTGA
Rentokil_     TTGTTATATCTGTGGTTGAATGGTTTTGGCACGCTTGTTCCTAACAATGCTTGTCATTGA
Scott_        TTGTTATATCTGTGGTTGAATTGTTTTGGCA---------------GTGCTTGTCATTGA
White_        TTGTTATATCTGTGGTTGAATGGTTTTGGCACGCTTGTTCCTAACAATGCTTGTCATTGA
Zurich_       TTGTTATATCTGTGGTTGAATGGTTTTGGCACGCTTGTTCCTAACAATGCTTGTCATTGA
              ********************** *********                 *************

Rutgers_      CTTTTAAGACTCCAAGGATCTTCTATTCGGCTAATATTAATCTTCAGATATAAATTCAAG
Cardiff_      CTTTTAAGACTCCAAGGATCTTCTATTCGGCTAATATTAATCTTCAGATATAAATTCAAG
Millan_       CTTTTAAGAC-CCAAGGATCTTCTATTCGGCTAATATTAATCTTCAAAATTAAATTCAAG
Rentokil_     CTTTTAAGAC-CCAAGGATCTTCTATTCGGCTAATATTAATCTTCAAAATTAAATTCAAG
Scott_        CTTTTAAGACTCCAAGGATCTTCTATTCGGCTAATATTAATCTTCAGATATAAATTCAAG
White_        CTTTTAAGAC-CCAAGGATCTTCTATTCGGCTAATATTAATCTTCAAAATTAAATTCAAG
Zurich_       CTTTTAAGAC-CCAAGGATCTTCTATTCGGCTAATATTAATCTTCAAAATTAAATTCAAG
              ********** ************************************ *  **********

Rutgers_      ATAAGATTCACAAAGATTTTGTTAATTTTATATGCGGACTCATA---TTTTTTTT-AATT
Cardiff_      ATAAGATTCACAAAGATTTTGTTAATTTTATATGCAGACTCATACAATTTTTT---AATT
Millan_       ATAAGATTCACAAAGATTTTGTTAATTTTACATGCGGACTCATACAATTTTTTTT-AATT
Rentokil_     ATAAGATTCACAAAGATTTTGTTAATTTTACATGCGGACTCATACAATTTTTTTT-AATT
Scott_        ATAAGATTCACAAAGATTTTGTTAATTTTATATGCAGACTCATACAATTTTTT---AATT
White_        ATAAGATTCACAAAGATTTTGTTAATTTTACATGCGGACTCATACAATTTTTTTTTAATT
Zurich_       ATAAGATTCACAAAGATTTTGTTAATTTTACATGCGGACTCATACAATTTTTTTTTAATT
              ******************************** **** ********   ******   ****

Rutgers_      TATTTTATGCCACTCTGCCGTGTTTTGGTCAAAAAGTTTGTAAAAACTCGAAATGATGGT
Cardiff_      TATTTTATGCCACTCTGCCGTGTTTTGGTCAAAAAGTTTGTAAAAACTCGAAATGATGGT
Millan_       TATTTTATGCCACTCTGCCGTGTTTTTGTCAAAAAGTTTGTAAAAACTCGAAATGATGGT
Rentokil_     TATTTTATGCCACTCTGCCGTGTTTTTGTCAAAAAGTTTGTAAAAACTCGAAATGATGGT
Scott_        TATTTTATGCCACTCTGCCGTGTTTTGGTCAAAAAGTTTGTAAAAACTCGAAATGATGGT
White_        TATTTTATGCCACTCTGCCGTGTTTTTGTCAAAAAGTTTGTAAAAACTCGAAATGATGGT
Zurich_       TATTTTATGCCACTCTGCCGTGTTTTTGTCAAAAAGTTTGTAAAAACTCGAAATGATGGT
              ************************** ***********************************

Rutgers_      CCTAGATCCACCATATTGTGTACCAATTAATATATCTCGACGACCAGTCCATTTTAAACT
Cardiff_      CCTAGATCCACCATATTGTGTACCAATTAATATATCTCGACGACCAGTCCATTTTAAACT
Millan_       CCTAGATCCACCATATTGTGTACCAATTAATATATCTCGACGACCAGTCCATTTTAAATT
```


Rentokil_ CCTAGATCCACCATATTGTGTACCAATTAATATATCTCGACGACCAGTCCATTTTAAATT
Scott_ CCTAGATCCACCATATTGTGTACCAATTAATATATCTCGACGACCAGTCCATTTTAAACT
White_ CCTAGATCCACCATATTGTGTACCAATTAATATATCTCGACGACCAGTCCATGTTAAACT
Zurich_ CCTAGATCCACCATATTGTGTACCAATTAATATATCTCGACGACCAGTCCATGTTAAACT
Rutgers_ GTAATTACTCCACAGGTCGCCGAATACAATCTTATAATGTTTGCAGATTATAGCGAATTT
Cardiff_ GTAATTACTCCACAGGTCGCCGAATACAATCTTATAATGTTTGCAGATTATAGCGAATTT
Millan_ GTAAATACTTCACAGGTCGCAGAATACAATCTTATGATATTTGCAGATGATAGTGAATTT
Rentokil_ GTAAATACTTCACAGGTCGCAGAATACAATCTTATGATATTTGCAGATGATAGTGAATTT
Scott_ GTAATTACTCCACAGGTCGCCGAATACAATCTTATAATGTTTGCAGATTATAGCGAATTT
White_ GTAATTACTCCACATGTCGCCGAATACAATCTTATAATGTTTGCAGATTATAGCGAATTT
Zurich_ GTAATTACTCCACATGTCGCCGAATACAATCTTATAATGTTTGCAGATTATAGCGAATTT
Rutgers_ TAAGATGGAGTTCATAATGTTGAAATATTTCTCATCAAAATTCACTATTATTATCAT--T
Cardiff_ TAAAATGGAGTTCATAATGTTAAAATATTTCTCATCAAAATCCACTATT---------TT
Millan_ AGAAACTGAGTTCTTAATGTTGAAATATTTCTCATCAAAATTCACTATTATTATCAT--T
Rentokil_ AGAAACTGAGTTCTTAATGTTGAAATATTTCTCATCAAAATTCACTATTATTATCAT--T
Scott_ TAAGATGGAGTTCATAATGTTAAAATATTTCTCATCAAAATCCACTATT----------T
White_ TAAGATGGAGTTCATAATGTTGAAATATTTCTCATCAAAATTCACTATTATTATCTTTTT
Zurich_ TAAGATGGAGTTCATAATGTTGAAATATTTCTCATCAAAATTCACTATTATTATCTTTTT
Rutgers_ TTTTTCTGAAATATAATTTTTTGTCATATATTGCTTCCCGTTGCAAAACCTATTTACCAA
Cardiff_ TTTTTC-GAAATATAATTTTTTGTCATATATTGCTTCCCGTTGTAAAACCTATTTACCAA
Millan_ TTTTTCTGAAATATAATTTGTTGTCATATATTGCTTCTCATTGCAAAACCTATTTACCAA
Rentokil_ TTTTTCTGAAATATAATTTGTTGTCATATATTGCTTCTCATTGCAAAACCTATTTACCAA
Scott_ TTTTTC-GAAATATAATTTTTTGTCATATATTGCTTCCCGTTGTAAAACCTATTTACCAA
White_ TTTTTCTGAAATATAATTTGTTGCCATATATTGCTTCTCGTTGTAAAACCTATTTTCCAA
Zurich_ TTTTTCTGAAATATAATTTGTTGCCATATATTGCTTCTCGTTGTAAAACCTATTTTCCAA
Rutgers_ TCATATATGATTTGTTTTTCTCAACTGAAACGAGGTTTTTCATAAA-AAAAC------CA
Cardiff_ TCATATATGATTT---TTTCTCAACTGAAACGAGGTTTTTCATAAACAAAAA------CA
Millan_ TCATATATGATTTT--TTTCTCAACTGAAACGAGGTTTTTCATAAAAAAAC-------CA
Rentokil_ TCATATATGATTTT--TTTCTCAACTGAAACGAGGTTTTTCATAAAAAAAC-------CA
Scott_ TCATATATGATTT---TTTCTCAACTGAAACGAGGTTTTTCATAAACAAAAA------CA
White_ TCATATATGATTTT--TT-CTCAACTAAAACGAGGTTTTTCATAAAAAAAAAAAAAA-CA
Zurich_ TCATATATGATTTT--TT-CTCAACTAAAACGAGGTTTTTCATAAAAAAAAAAAAAAACA
Rutgers_ AAATCTTTGCAAAGTTCAATTGCTTTCCCTTTAATCCCAATAATTACATCAACACGCTTT
Cardiff_ AAATCTTTGCAAAGTTCAATTGTTTTCCCTTTAATCCCAATAATTACATCAACACGCTTT
Millan_ AAATCTTTGCAAAGTTCAATTGCTTTCCCTTTAATCCCAATAATTACATCAACACGCTTT
Rentokil_ AAATCTTTGCAAAGTTCAATTGCTTTCCCTTTAATCCCAATAATTACATCAACACGCTTT
Scott_ AAATCTTTGCAAAGTTCAATTGCTTTCCCTTTAATCCCAATAATTACATCAACACGCTTT
White_ AAATCTTTGCAAAGTTCAATTGCTTTCCCTTTAATCCCAATAATTACATCAACACGCTTT
Zurich_ AAATCTTTGCAAAGTTCAATTGCTTTCCCTTTAATCCCAATAATTACATCAACACGCTTT
Rutgers_ GCCCCCAATACAAAATGAGAAACATTAAGAATTCCAATGTTTGATTTTTTCAATAGTTCA
Cardiff_ GCCCCCAATACAAAATGAGAAACATTAAGAATTCCAATGTTTGATTTTTTCAATAGTTCA
Millan_ GCCCCCAGTACAAAATGAGAAACATTAAGAATTCCAATGTTTGATTTTTTCAATAGTTCA
Rentokil_ GCCCCCAGTACAAAATGAGAAACATTAAGAATTCCAATGTTTGATTTTTTCAATAGTTCA
Scott_ GCCCCCAATACAAAATGAGAAACATTAAGAATTCCAGTGTTTGATTTTTTCAATAGTTCA
White_ GCCCCCAGTACAAAATGAGAAACATTAAGAATTCCAGTGTTTGATTTTTTCAATAGTTCA
Zurich_ GCCCCCAATACAAAATGAGAAACATTAAGAATTCCAGTGTTTGATTTTTTCAATAGTTCA

```
                ******* **************************** ***********************

Rutgers_        GGGATTATCCAAAATAAAGAATAAAATTTAGAATTTCCCAACTGCATATCAAACAGCCAT
Cardiff_        GGGATTATCCAAAATAAAGAATAAAATTTAGAATTTCCCAACTGCATATCAAACAGCCAT
Millan_         GGGATTATCCAAAATAAAGAATAAAATTTAGAAATTCCCAACTGCATATCAAACAGCCAT
Rentokil_       GGGATTATCCAAAATAAAGAATAAAATTTAGAAATTCCCAACTGCATATCAAACAGCCAT
Scott_          GGGATTATCCAAAATAAAGAATAAAATTTAGAAATTCCCAACTGCATATCAAACAGCCAT
White_          GGGATTATCCAAAATAAAGAATAAAATTTAGAAATTCCCAACTGCATATCAAACAGCCAT
Zurich_         GGGATTATCCAAAATAAAGAATAAAATTTAGAAATTCCCAACTGCATATCAAACAGCCAT
                ********************************** ************************

Rutgers_        CGACAATTAATTCCAAGTAAACGAAAAAA---CTATATTCCTCTTGACAATGAATGCAAA
Cardiff_        CGACAATTAATTCCAAGTAAACGAAAAAAA--CTATATTCCTCTTGACAATGAATGCAAA
Millan_         CGACAATTAATTCCAAGTAAACGAAAAAAAA-CTATATTCCTCTTGACAATGAATGCAAA
Rentokil_       CGACAATTAATTCCAAGTAAACGAAAAAAAA-CTATATTCCTCTTGACAATGAATGCAAA
Scott_          CGACAATTAATTCCAAGTAAACGAAAAAA---CTATATTCCTCTTGACAATGAATGCAAA
White_          CGACAATTAATTCCAAGTAAACGAAAAAAAAACTATATTCCTCTTGACAATGAATGCAAA
Zurich_         CGACAATTAATTCCAAGTAAACGAAAAAAAAACTATATTCCTCTTGACAATGAATGCAAA
                ****************************   ****************************

Rutgers_        AAGAGAGCAATTAAGTGAAGCAAAGAAAGAAAACATCATTAGAGAAAATTAAAACGTTAG
Cardiff_        AAGAGAGCAATTAAGTGAAGCAAAGAAAGAAAACATCATTAGAGAAAATTAAAACGTTAG
Millan_         AAGAGAGCAATTAAGTGAACCAAAGAAAGAAAACATCATTAGAGAAAATTAAAACGTTAG
Rentokil_       AAGAGAGCAATTAAGTGAACCAAAGAAAGAAAACATCATTAGAGAAAATTAAAACGTTAG
Scott_          AAGAGAGCAATTAAGTGAAGCAAAGAAAGAAAACATCATTAGAGAAAATTAAAACGTTAG
White_          AAGAGAGCAATTAAGTGAACCAAAGAAAGAAAACATCATTAGAGAAAATTAAAACGTTAG
Zurich_         AAGAGAGCAATTAAGTGAACCAAAGAAAGAAAACATCATTAGAGAAAATTAAAACGTTAG
                ******************** ****************************************

Rutgers_        TAGAACATCTAGAGCACCGGAGAACAGGTAACAACAACGATGTACATTAACAATGGTCCT
Cardiff_        TAGAACATCTAGAGCACCGGAGAACAGGTAACAACAACGATGTACATTAACAATGGTCCT
Millan_         TAGAACATCTAGAGCACCGGAGAACAGGTAACAACAACGATGTACATTAACAATGGTCCT
Rentokil_       TAGAACATCTAGAGCACCGGAGAACAGGTAACAACAACGATGTACATTAACAATGGTCCT
Scott_          TAGAACATCTAGAGCACCGGAGAACAGGTAACAACAACGATGTACATTAACAATGGTCCT
White_          TAGAACATCTAGAGCACCGGAGAACAGGTAACAACAACGATGTACATTAACAATGGTCCT
Zurich_         TAGAACATCTAGAGCACCGGAGAACAGGTAACAACAACGATGTACATTAACAATGGTCCT
                ************************************************************

Rutgers_        TCAGTTCAGTTGATGGAAGAGCAGCGCACACAGCACAGTGGCAATAAACGTGCCACCCTT
Cardiff_        TCAGTTCAGTTGATGGAAGAGCAGCGCACACAGCACAGTGGCAATAAACGTGCCACCCTT
Millan_         TCAGTTCAGTTGATGGAAGAGCAGCGCACACAGCACAGTGGCAATAAACGTGCCACCCTT
Rentokil_       TCAGTTCAGTTGATGGAAGAGCAGCGCACACAGCACAGTGGCAATAAACGTGCCACCCTT
Scott_          TCAGTTCAGTTGATGGAAGAGCAGCGCACACAGCACAGTGGCAATAAACGTGCCACCCTT
White_          TCAGTTCAGTTGATGGAAGAGCAGCGCACACAGCACAGTGGCAATAAACGTGCCACCCTT
Zurich_         TCAGTTCAGTTGATGGAAGAGCAGCGCACACAGCACAGTGGCAATAAACGTGCCACCCTT
                ************************************************************

Rutgers_        TTCTCACAGGCACCGGGTACAATAATTCTTTCATAAAATTTTCGAAATGAGAGCAGAACA
Cardiff_        TTCTCACAGGCACCGGGTACAATAATTCTTTCATAAAATTTTCGAAATGAGAGCAGAACA
Millan_         TTCTCACAGGCACCGGGTACAATAATTCTTTCATAAAATTTTCGAAATGAGAGCAGAACA
Rentokil_       TTCTCACAGGCACCGGGTACAATAATTCTTTCATAAAATTTTCGAAATGAGAGCAGAACA
Scott_          TTCTCACAGGCACCGGGTACAATAATTCTTTCATAAAATTTTCGAAATGAGAGCAGAACA
White_          TTCTCACAGGCACCGGGTACAATAATTCTTTCATAAAATTTTCGAAATGAGAGCAGAACA
Zurich_         TTCTCACAGGCACCGGGTACAATAATTCTTTCATAAAATTTTCGAAATGAGAGCAGAACA
                ************************************************************

Rutgers_        CAGGTAAACCACCATAAAAGAGATGTCATAAAATCATCACCGGCACTGAATGAAAACCTC
Cardiff_        CAGGTAAACCACCATAAAAGAGATGTCATAAAATCATCACCGGCACTGAATGAAAACCTC
```

```
Millan_          CAGGTAAACCACCATAAAAGAGATGTCATAAAATCATCACCGGCACTGAATGAAAACCTC
Rentokil_        CAGGTAAACCACCATAAAAGAGATGTCATAAAATCATCACCGGCACTGAATGAAAACCTC
Scott_           CAGGTAAACCACCATAAAAGAGATGTCATAAAATCATCACCGGCACTGAATGAAAACCTC
White_           CAGGTAAACCACCATAAAAGAGATGTCATAAAATCATCACCGGCACTGAATGAAAACCTC
Zurich_          CAGGTAAACCACCATAAAAGAGATGTCATAAAATCATCACCGGCACTGAATGAAAACCTC
                 ************************************************************

Rutgers_         TGCCAATGTATGCGAATATACACACGCACAGGCCAAAAGATCTGCACGAGGGGTTGTTCC
Cardiff_         TGCCAATGTATGCGAATATACACACGCACAGGCCAAAAGATCTGCACAAGGGGTTGTTCC
Millan_          TGCCAATGTATGCGAATATACACACGCACAGGCCAAAAGATCTGCACAAGGGGTTGTTCC
Rentokil_        TGCCAATGTATGCGAATATACACACGCACAGGCCAAAAGATCTGCACAAGGGGTTGTTCC
Scott_           TGCCAATGTATGCGAATATACACACGCACAGGCCAAAAGATCTGCACGAGGGGTTGTTCC
White_           TGCCAATGTATGCGAATATACACACGCACAGGCCAAAAGATCTGCACAAGGGGTTGTTCC
Zurich_          TGCCAATGTATGCGAATATACACACGCACAGGCCAAAAGATCTGCACAAGGGGTTGTTCC
                 ********************************************** ************

Rutgers_         ACACACACGTACACGAATGCTCTTCATTAAAGGATAATGTGAGAGAAAGAGAACAAAAAT
Cardiff_         ACACACACGTACACGAATGCTCTTCATTAAAGGATAATGTGAGAGAAAGAGAACAAAAAT
Millan_          ACACACACGTACACGAATGCTCTTCATTAAAGGATAATGTGAGAGAAAGAGAACAAAAAT
Rentokil_        ACACACACGTACACGAATGCTCTTCATTAAAGGATAATGTGAGAGAAAGAGAACAAAAAT
Scott_           ACACACACGTACACGAATGCTCTTCATTAAAGGATAATGTGAGAGAAAGAGAACAAAAAT
White_           ACACACACGTACACGAATGCTCTTCATTAAAGGATAATGTGAGAGAAAGAGAACAAAAAT
Zurich_          ACACACACGTACACGAATGCTCTTCATTAAAGGATAATGTGAGAGAAAGAGAACAAAAAT
                 ************************************************************

Rutgers_         GCACCCTCTTGGCTGTGTGTGTAACCAATGAACGAACGCTTACTAGGTAAAAGGGGATGA
Cardiff_         GCACCCTCTTGGCTGTGTGTGTAACCAATGAACGAACGCTTACTAGGTAAAAGGGGATGA
Millan_          GCACCCTCTTGGCTGTGTGTGTAACCAATGAACGAACGCTTACTAGGTAAAAGGGGATGA
Rentokil_        GCACCCTCTTGGCTGTGTGTGTAACCAATGAACGAACGCTTACTAGGTAAAAGGGGATGA
Scott_           GCACCCTCTTGGCTGTGTGTGTAACCAATGAACGAACGCTTACTAGGTAAAAGGGGATGA
White_           GCACCCTCTTGGCTGTGTGTGTAACCAATGAACGAACGCTTACTAGGTAAAAGGGGATGA
Zurich_          GCACCCTCTTGGCTGTGTGTGTAACCAATGAACGAACGCTTACTAGGTAAAAGGGGATGA
                 ************************************************************

Rutgers_         TCGTAACATTGGTTGGCCGTATGAATGGTAGAGTGCAAAATTTGCGCCAGGACTA
Cardiff_         TCGTAACATTGGTTGGCCGTATGAATGGTAGAGTGCAAAATTTGCGCCAGGACTA
Millan_          TCGTAACATTGGTTGGCCGTATGAATGGTAGAGTGCAAAATTTGCGCCAGGACTA
Rentokil_        TCGTAACATTGGTTGGCCGTATGAATGGTAGAGTGCAAAATTTGCGCCAGGACTA
Scott_           TCGTAACATTGGTTGGCCGTATGAATGGTAGAGTGCAAAATTTGCGCCAGGACTA
White_           TCGTAACATTGGTTGGCCGTATGAATGGTAGAGTGCAAAATTTGCGCCAGGACTA
Zurich_          TCGTAACATTGGTTGGCCGTATGAATGGTAGAGTGCAAAATTTGCGCCAGGACTA
                 ************************************************************
```

tll 5'UTR alignment

```
Rutgers_         TTAAAGGACGAGGACGGTGGTGAATTAGCAACACAAACTATTTGGATCTCAAACAGTGAA
Cardiff_         TTAAAGGACGAGGACGGTGGTGAATTAGCAACACAAACTATTTGGATCTCAAACAGTGAA
Millan_          TTAAAGGACGAGGACGGTGGTGAATTAGCAACACAAACTATTTGGATCTCAAACAGTGAA
Rentokil_        TTAAAGGACGAGGACGGTGGTGAATTAGCAACACAAACTATTTGGATCTCAAACAGTGAA
Scott_           TTAAAGGACGAGGACGGTGGTGAATTAGCAACACAAACTATTTGGATCTCAAACAGTGAA
White_           TTAAAGGACGAGGACGGTGGTGAATTAGCAACACAAACTATTTGGATCTCAAACAGTGAA
Zurich_          TTAAAGGACGAGGACGGTGGTGAATTAGCAACACAAACTATTTGGATCTCAAACAGTGAA
                 ************************************************************

Rutgers_         CACAACTCAAACAGAGCTGAGAACACTAAAAATAAAACAAAATATCTTTACAACAATTAC
Cardiff_         CACAACTCAAACAGAGCTGAGAACACTAAAAATAAAACAAAATATCTTTACAACAATTAC
Millan_          CACAACTCAAACAGAGCTGAGAACACTAAAA-TAAAACAAAATATATTTACAACAATTAC
Rentokil_        CACAACTCAAACAGAGCTGAGAACACTAAAAATAAAACAAAATATCTTTACAACAATTAC
```

```
Scott_          CACAACTCAAACAGAGCTGAGAACACTAAAAATAAAACAAAATATCTTTACAACAATTAC
White_          CACAACTCAAACAGAGCTGAGAACACTAAAAATAAAACAAAATATCTTTACAACAATTAC
Zurich_         CACAACTCAAACAGAGCTGAGAACACTAAAAATTAA-CAAAATATCTTTACAACAATTAC
                ********************************* * ** ******** **************

Rutgers_        GAATTAAAAAATATTTGATATAACAAAAAACATACATTTAACCACGATCAAGGATTACTT
Cardiff_        GAATTAAAAAATATTTGATATAACAAAAAACATACATTTAACCACGATCAAGGATTACTT
Millan_         GAATTAAAAA-TATTTGATAGAACAAAAAACATACATTTAACCACGATCAAGGATTACTT
Rentokil_       GAATTAAAAAATATTTGATATAACCAAAAACATACATTTAACCACTATCAAGGATTACTT
Scott_          GAATTAAAAAATATTTGATATAACAAAAAACATACATTTAACCACGATCAAGGATTACTT
White_          GAATTAAAAAATATTTGATATAACCAAAAACATACATTTAACCACGATCAAGGATTACTT
Zurich_         GAATTAAAAAATATTTGATATAACCAAAAACATACATTTAACCACGATCAAGGATTACTT
                ********** ********* *** ********************* **************

Rutgers_        TACAATAACAAACAACAAA---
Cardiff_        TACAATAACAAACAACAAA---
Millan_         TACAATAACAAACAACAAA---
Rentokil_       TACAATAACAAACAACAAA---
Scott_          TACAATAACAGAAAACAACAAA
White_          TACAATAACAGAAAACAACAAA
Zurich_         TACAATAACAAACAACAAA---
                ********** * *****
```

tll coding sequence alignment

```
Rutgers_        ATGCAAACCACCGAAGGATCTCCCGATATTATGGATCAAAAATACAACTCCGTCAGATTA
Cardiff_        ATGCAAACCACCGAAGGATCTCCCGATATTATGGATCAAAAATACAACTCCGTCAGATTA
Millan_         ATGCAAACCACCGAAGGATCTCCCGATATTATGGATCAAAAATACAACTCCGTCAGATTA
Rentokil_       ATGCAAACCACCGAAGGATCTCCCGATATTATGGATCAAAAATACAACTCCGTCAGATTA
Scott_          ATGCAAACCACCGAAGGATCTCCCGATATTATGGATCAAAAATACAACTCCGTCAGATTA
White_          ATGCAAACCACCGAAGGATCTCCCGATATTATGGATCAAAAATACAACTCCGTCAGATTA
Zurich_         ATGCAAACCACCGAAGGATCTCCCGATATTATGGATCAAAAATACAACTCCGTCAGATTA
                ************************************************************

Rutgers_        TCTCCAGCTGCGTCAAGTCGCATCCTATATCATGTGCCTTGCAAAGTTTGCCGTGACCAC
Cardiff_        TCTCCAGCTGCGTCAAGTCGCATCCTATATCATGTGCCTTGCAAAGTTTGCCGTGACCAC
Millan_         TCTCCAGCTGCGTCAAGTCGCATCCTATATCATGTGCCTTGCAAAGTTTGCCGTGACCAT
Rentokil_       TCTCCAGCTGCGTCAAGTCGCATCCTATATCATGTGCCTTGCAAAGTTTGCCGTGACCAT
Scott_          TCTCCAGCTGCGTCAAGTCGCATCCTATATCATGTGCCTTGCAAAGTTTGCCGTGACCAC
White_          TCTCCAGCTGCGTCAAGTCGCATCCTATATCATGTGCCTTGCAAAGTTTGCCGTGACCAC
Zurich_         TCTCCAGCTGCGTCAAGTCGCATCCTATATCATGTGCCTTGCAAAGTTTGCCGTGACCAC
                ************************************************************

Rutgers_        AGTTCTGGCAAACATTACGGTATCTATGCATGTGATGGCTGTGCTGGTTTCTTCAAGCGT
Cardiff_        AGTTCTGGCAAACATTACGGTATCTATGCATGTGATGGCTGTGCTGGTTTCTTTAAGCGT
Millan_         AGTTCTGGCAAACATTACGGTATCTATGCATGTGATGGCTGTGCTGGTTTCTTTAAGCGT
Rentokil_       AGTTCTGGCAAACATTACGGTATCTATGCATGTGATGGCTGTGCTGGTTTCTTTAAGCGT
Scott_          AGTTCTGGCAAACATTACGGTATCTATGCATGTGATGGCTGTGCTGGTTTCTTTAAGCGT
White_          AGTTCTGGCAAACATTACGGTATCTATGCATGTGATGGCTGTGCTGGTTTCTTCAAGCGT
Zurich_         AGTTCTGGCAAACATTACGGTATCTATGCATGTGATGGCTGTGCTGGTTTCTTTAAGCGT
                ************************************************** ******

Rutgers_        TCCATTCGGCGTTCCCGCCAATATGTGTGCAAATCCCAGAAACAAGGACTCTGTGTGGTG
Cardiff_        TCCATTCGGCGTTCCCGCCAATATGTGTGCAAATCCCAGAAACAAGGACTCTGTGTGGTG
Millan_         TCCATTCGGCGTTCCCGCCAATATGTGTGCAAATCCCAGAAACAAGGACTCTGTGTGGTG
Rentokil_       TCCATTCGGCGTTCCCGCCAATATGTGTGCAAATCCCAGAAACAAGGACTCTGTGTGGTG
Scott_          TCCATTCGGCGTTCCCGCCAATATGTGTGCAAATCCCAGAAACAAGGACTCTGTGTGGTG
White_          TCCATTCGGCGTTCCCGCCAATATGTGTGCAAATCCCAGAAACAAGGACTCTGTGTGGTG
```

```
Zurich_         TCCATTCGGCGTTCCCGCCAATATGTGTGCAAATCCCAGAAACAAGGACTCTGTGTGGTG
                ************************************************************

Rutgers_        GACAAAACCCATCGCAATCAGTGCCGTGCCTGCCGCTTGCGCAAATGTTTCGAAGTTGGC
Cardiff_        GACAAAACCCATCGCAATCAGTGCCGTGCCTGCCGCTTGCGCAAATGTTTCGAAGTTGGC
Millan_         GACAAAACCCATCGCAATCAGTGCCGTGCCTGCCGCTTGCGCAAATGTTTCGAAGTTGGC
Rentokil_       GACAAAACCCATCGCAATCAGTGCCGTGCCTGCCGCTTGCGCAAATGTTTCGAAGTTGGC
Scott_          GACAAAACCCATCGCAATCAGTGCCGTGCCTGCCGCTTGCGCAAATGTTTCGAAGTTGGC
White_          GACAAAACCCATCGCAATCAGTGCCGTGCCTGCCGCTTGCGCAAATGTTTCGAAGTTGGC
Zurich_         GACAAAACCCATCGCAATCAGTGCCGTGCCTGCCGCTTGCGCAAATGTTTCGAAGTTGGC
                ************************************************************

Rutgers_        ATGAACAAAGATGCTGTCCAACACGAACGTGGACCCCGCAACTCCACATTGCGCCGCCAC
Cardiff_        ATGAACAAAGATGCTGTCCAACACGAACGTGGACCCCGCAACTCTACATTGCGCCGCCAC
Millan_         ATGAACAAAGATGCTGTCCAACACGAACGTGGACCCCGCAACTCCACATTGCGCCGCCAC
Rentokil_       ATGAACAAAGATGCTGTCCAACACGAACGTGGACCCCGCAACTCCACATTGCGCCGCCAC
Scott_          ATGAACAAAGATGCTGTCCAACACGAACGTGGACCCCGCAACTCTACATTGCGCCGCCAC
White_          ATGAACAAAGATGCTGTCCAACACGAACGTGGACCCCGCAACTCCACATTGCGCCGCCAC
Zurich_         ATGAACAAAGATGCTGTCCAACACGAACGTGGACCCCGCAACTCTACATTGCGCCGCCAC
                ******************************************** ***************

Rutgers_        ATGGCAATGTACAAAGATGCCATGATGGGTGGCTCTGAAATGCCCCAGATCCCAGCTGAA
Cardiff_        ATGGCCATGTACAAAGATGCCATGATGGGTGGCTCTGAAATGCCCCAGATCCCAGCTGAA
Millan_         ATGGCCATGTACAAAGATGCCATGATGGGTGGCTCTGAAATGCCCCAGATCCCAGCTGAA
Rentokil_       ATGGCCATGTACAAAGATGCCATGATGGGTGGCTCTGAAATGCCCCAGATCCCAGCTGAA
Scott_          ATGGCCATGTACAAAGATGCCATGATGGGTGGCTCTGAAATGCCCCAGATCCCAGCTGAA
White_          ATGGCAATGTACAAAGATGCCATGATGGGTGGCTCTGAAATGCCCCAGATCCCAGCTGAA
Zurich_         ATGGCCATGTACAAAGATGCCATGATGGGTGGCTCTGAAATGCCCCAGATCCCAGCTGAA
                ***** ********************************************************

Rutgers_        ATTCTCATGAACACCGCAGCTTTAACTGGTTTCCCCGGCTTGCCAATGCCAATTCCAGGA
Cardiff_        ATTCTCATGAACACCGCAGCTTTAACTGGTTTCCCCGGCTTGCCAATGCCAATTCCAGGA
Millan_         ATTCTTATGAACACCGCAGCTTTAACTGGTTTCCCCGGCTTGCCAATGCCAATTCCAGGA
Rentokil_       ATTCTTATGAACACCGCAGCTTTAACTGGTTTCCCCGGCTTGCCAATGCCAATTCCAGGA
Scott_          ATTCTCATGAACACCGCAGCTTTAACTGGTTTCCCCGGCTTGCCAATGCCAATTCCAGGA
White_          ATTCTTATGAACACCGCAGCTTTAACTGGTTTCCCCGGCTTGCCAATGCCAATTCCAGGA
Zurich_         ATTCTTATGAACACCGCAGCTTTAACTGGTTTCCCCGGCTTGCCAATGCCAATTCCAGGA
                ***** ********************************************************

Rutgers_        TCCCATCACATGCATCCCAGTTTGGCTGGAGCCTTCCCTGCACCACCATCAGTTTTGGAT
Cardiff_        TCCCATCACATGCATCCCAGTTTGGCTGGAGCCTTCCCTGCACCACCATCAGTTTTGGAT
Millan_         TCCCATCATATGCATCCCAGTTTGGCTGGAGCCTTCCCTGCACCACCATCAGTTTTGGAT
Rentokil_       TCCCATCATATGCATCCCAGTTTGGCTGGAGCCTTCCCTGCACCACCATCAGTTTTGGAT
Scott_          TCCCATCATATGCATCCCAGTTTGGCTGGAGCCTTCCCTGCACCACCATCAGTTTTGGAT
White_          TCCCATCATATGCATCCCAGTTTGGCTGGAGCCTTCCCTGCACCACCATCAGTTTTGGAT
Zurich_         TCCCATCACATGCATCCCAGTTTGGCTGGAGCCTTCCCTGCACCACCATCAGTTTTGGAT
                ******** ***************************************************

Rutgers_        TTATCTGTACCTCGTGTACCCCAACATCCCATGCATCAAGCTCATCCCGGTTTCTTTGCA
Cardiff_        TTATCTGTACCTCGTGTACCCCAACATCCCATGCATCAAGCTCATCCCGGTTTCTTTGCA
Millan_         TTATCTGTACCTCGTGTACCCCAACATCCCATGCATCAAGCTCATCCCGGTTTCTTTGCA
Rentokil_       TTATCTGTACCTCGTGTACCCCAACATCCCATGCATCAAGCTCATCCCGGTTTCTTTGCA
Scott_          TTATCTGTACCTCGTGTACCCCAACATCCCATGCATCAAGCTCATCCCGGTTTCTTTGCA
White_          TTATCTGTACCTCGTGTACCCCAACATCCCATGCATCAAGCTCATCCCGGTTTCTTTGCA
Zurich_         TTATCTGTACCTCGTGTACCCCAACATCCCATGCATCAAGCTCATCCCGGTTTCTTTGCA
                ************************************************************

Rutgers_        CCCACTGCCGCCTACATGAATGCCTTGGCTGCCACCCGTGTCTTGCCACCCACACCTCCA
```

```
Cardiff_        CCCACTGCCGCCTACATGAATGCCTTGGCTGCCACCCGTGTCTTGCCACCCACACCTCCA
Millan_         CCCACTGCCGCCTACATGAATGCCTTGGCTGCCACCCGTGTCTTGCCACCCACACCTCCA
Rentokil_       CCCACTGCCGCCTACATGAATGCCTTGGCTGCCACCCGTGTCTTGCCACCCACACCTCCA
Scott_          CCCACTGCCGCCTACATGAATGCCTTGGCTGCCACCCGTGTCTTGCCACCCACACCTCCA
White_          CCCACTGCCGCCTACATGAATGCCTTGGCTGCCACCCGTGTCTTGCCACCCACACCTCCA
Zurich_         CCCACTGCCGCCTACATGAATGCCTTGGCTGCCACCCGTGTCTTGCCACCCACACCTCCA
                ************************************************************

Rutgers_        CTAATGGCTGCCGAACACATTAAAGAAACTGCAGCCGAACATCTCTTCAAGAACATCAAC
Cardiff_        CTAATGGCTGCCGAACACATTAAAGAAACTGCCGCTGAACATCTCTTCAAGAACATCAAC
Millan_         CTAATGGCTGCCGAACACATTAAAGAAACTGCCGCCGAACATCTCTTCAAGAACATCAAC
Rentokil_       CTAATGGCTGCCGAACACATTAAAGAAACTGCCGCCGAACATCTCTTCAAGAACATCAAC
Scott_          CTAATGGCTGCCGAACACATTAAAGAAACTGCCGCCGAACATCTCTTCAAGAACATCAAC
White_          CTAATGGCTGCCGAACACATTAAAGAAACTGCCGCCGAACATCTCTTCAAGAACATCAAC
Zurich_         CTAATGGCTGCCGAACACATTAAAGAAACTGCCGCTGAACATCTCTTCAAGAACATCAAC
                ********************************* ** ************************

Rutgers_        TGGATTAAGAACGTACCCTCATTTGGTGAATTGCCACTGCCCGATCAATTGCAGTTGTTG
Cardiff_        TGGATTAAGAACGTGCCCTCATTCGGTGAATTGCCACTGCCCGATCAATTGCAGTTGTTG
Millan_         TGGATTAAGAACGTGCCCTCATTCGGTGAATTGCCTCTGCCCGATCAATTGCAGTTGTTG
Rentokil_       TGGATTAAGAACGTGCCCTCATTCGGTGAATTGCCTCTGCCCGATCAATTGCAGTTGTTG
Scott_          TGGATTAAGAACGTGCCCTCATTCGGTGAATTGCCACTGCCCGATCAATTGCAGTTGTTG
White_          TGGATTAAGAACGTGCCCTCATTCGGTGAATTGCCTCTGCCCGATCAATTGCAGTTGTTG
Zurich_         TGGATTAAGAACGTGCCCTCATTCGGTGAATTGCCACTGCCCGATCAATTGCAGTTGTTG
                ************** ******** *********** ************************

Rutgers_        GAAGACTCCTGGAAGGAATTCTTCATCTTGGCTATGGCGCAATACCTCATGCCCATGAAC
Cardiff_        GAAGACTCCTGGAAGGAATTCTTCATCTTGGCTATGGCGCAATACCTCATGCCCATGAAC
Millan_         GAAGACTCCTGGAAGGAATTCTTCATCTTGGCCATGGCGCAATATCTCATGCCCATGAAC
Rentokil_       GAAGACTCCTGGAAGGAATTCTTCATCTTGGCCATGGCGCAATATCTCATGCCCATGAAC
Scott_          GAAGACTCCTGGAAGGAATTCTTCATCTTGGCCATGGCGCAATATCTCATGCCCATGAAC
White_          GAAGACTCCTGGAAGGAATTCTTCATCTTGGCCATGGCGCAATATCTCATGCCCATGAAC
Zurich_         GAAGACTCCTGGAAGGAATTCTTCATCTTGGCCATGGCGCAATACCTCATGCCCATGAAC
                ********************************* *********** ***************

Rutgers_        TTCACTCAGCTATTGTTCGTTTACGAATCGGAAAATCCCAACCGCGATGTTACCGGTTTG
Cardiff_        TTCACTCAGCTATTGTTCGTTTACGAATCGGAAAATCCCAACCGCGATGTTACCGGTTTG
Millan_         TTCACTCAGCTATTGTTCGTTTACGAATCGGAAAATCCCAACCGCGATGTTACCGGTTTG
Rentokil_       TTCACTCAGCTATTGTTCGTTTACGAATCGGAAAATCCCAACCGCGATGTTACCGGTTTG
Scott_          TTCACTCAGCTATTGTTCGTTTACGAATCGGAAAATCCCAACCGCGATGTTACCGGTTTG
White_          TTCACTCAGCTATTGTTCGTTTACGAATCGGAAAATCCCAACCGCGATGTTACCGGTTTG
Zurich_         TTCACTCAGCTATTGTTCGTTTACGAATCGGAAAATCCCAACCGCGATGTTACCGGTTTG
                ************************************************************

Rutgers_        GTAACCCGTGAAGTTCATGCCTTCCAAGATGTCCTAAATCAATTGTGTCATCTCAACATT
Cardiff_        GTAACCCGTGAAGTGCATGCCTTCCAAGATGTCCTAAATCAATTGTGTCATCTCAACATT
Millan_         GTAACCCGTGAAGTGCATGCCTTCCAAGATGTCCTAAATCAATTGTGTCATCTCAACATT
Rentokil_       GTAACCCGTGAAGTGCATGCCTTCCAAGATGTCCTAAATCAATTGTGTCATCTCAACATT
Scott_          GTAACCCGTGAAGTGCATGCCTTCCAAGATGTCCTAAATCAATTGTGTCATCTCAACATT
White_          GTAACCCGTGAAGTGCATGCCTTCCAAGATGTCCTAAATCAATTGTGTCATCTCAACATT
Zurich_         GTAACCCGTGAAGTTCATGCCTTCCAAGATGTCCTAAATCAATTGTGTCATCTCAACATT
                ************** ********************************************

Rutgers_        GATAGCCATGAATATGAGCTCATTCGTGCCTTGACCCTATTCCGCCGCCCTGGCTCCGAT
Cardiff_        GATAGCCATGAATATGAGCTCATTCGTGCCTTGACCCTATTCCGCCGCCCTGGCTCCGAT
Millan_         GATAGCCATGAATATGAGCTCATTCGTGCCTTGACCCTATTCCGCCGCCCTGGCTCCGAT
Rentokil_       GATAGCCATGAATATGAGCTCATTCGTGCCTTGACCCTATTCCGCCGCCCTGGCTCCGAT
Scott_          GATAGCCATGAATATGAGCTCATTCGTGCCTTGACCCTATTCCGCCGCCCTGGCTCCGAT
```

```
White_          GATAGCCATGAATATGAGCTCATTCGTGCCTTGACCCTATTCCGCCGCCCTGGCTCCGAT
Zurich_         GATAGCCATGAATATGAGCTCATTCGTGCCTTGACCCTATTCCGCCGCCCTGGCTCCGAT
                ************************************************************

Rutgers_        GATTTGGCCAATTCTTCACTGTCCACCAGCAACGGCAGTCCCAACTCCAGCATCTCTGCC
Cardiff_        GATTTGGCCAATTCTTCACTGTCCACCAGCAACGGCAGTCCCAACTCCAGCATCTCTGCT
Millan_         GATTTGGCCAATTCTTCACTGTCCACCAGCAACGGCAGTCCCAACTCCAGCATCTCTGCC
Rentokil_       GATTTGGCCAATTCTTCACTGTCCACCAGCAACGGCAGTCCCAACTCCAGCATCTCTGCC
Scott_          GATTTGGCCAATTCTTCACTGTCCACCAGCAACGGCAGTCCCAACTCCAGCATCTCTGCC
White_          GATTTGGCCAATTCTTCACTGTCCACCAGCAACGGCAGTCCCAACTCCAGCATCTCTGCC
Zurich_         GATTTGGCCAATTCTTCACTGTCCACCAGCAACGGCAGTCCCAACTCCAGCATCTCTGCC
                ************************************************************

Rutgers_        GAATCTCGTGGCCTCATCGAAAGCACCAAAATTGCCGCCTTACATGATGAGAGCCGTAAT
Cardiff_        GAATCTCGTGGCCTCATCGAAAGCACCAAAATTGCCGCCTTACATGATGAGAGCCGTAAT
Millan_         GAATCTCGTGGCCTCATCGAAAGCACCAAAATTGCCGCCTTACATGATGAGAGCCGTAAT
Rentokil_       GAATCTCGTGGCCTCATCGAAAGCACCAAAATTGCCGCCTTACATGATGAGAGCCGTAAT
Scott_          GAATCTCGTGGCCTCATCGAAAGCACCAAAATTGCCGCCTTACATGATGAGAGCCGTAAT
White_          GAATCTCGTGGCCTCATCGAAAGCACCAAAATTGCCGCCTTACATGATGAGAGCCGTAAT
Zurich_         GAATCTCGTGGCCTCATCGAAAGCACCAAAATTGCCGCCTTACATGATGAGAGCCGTAAT
                ************************************************************

Rutgers_        GCCCTCATTGGCTACATTGCCCGTCTACATCCCGGCCAACCCATGCGTTTCCAAAGCATT
Cardiff_        GCCCTCATTGGCTACATTGCCCGTCTACATCCCGGCCAACCCATGCGTTTCCAAAGCATT
Millan_         GCCCTCATTGGCTACATTGCCCGTCTACATCCCGGCCAACCCATGCGTTTCCAAAGCATT
Rentokil_       GCCCTCATTGGCTACATTGCCCGTCTACATCCCGGCCAACCCATGCGTTTCCAAAGCATT
Scott_          GCCCTCATTGGCTACATTGCCCGTCTACATCCCGGCCAACCCATGCGTTTCCAAAGCATT
White_          GCCCTCATTGGCTACATTGCCCGTCTACATCCCGGCCAACCCATGCGTTTCCAAAGCATT
Zurich_         GCCCTCATTGGCTACATTGCCCGTCTACATCCCGGCCAACCCATGCGTTTCCAAAGCATT
                ************************************************************

Rutgers_        ATGAGTGTATTGACCCAGATGCACAAAGTCTCCTCGTTTGCCATTGAGGAATTGTTCTTC
Cardiff_        ATGAGTGTATTGACCCAGATGCACAAAGTCTCCTCGTTTGCCATTGAGGAATTGTTCTTC
Millan_         ATGAGTGTATTGACCCAGATGCACAAAGTCTCCTCGTTTGCCATTGAGGAATTGTTCTTC
Rentokil_       ATGAGTGTATTGACCCAGATGCACAAAGTCTCCTCGTTTGCCATTGAGGAATTGTTCTTC
Scott_          ATGAGTGTATTGACCCAGATGCACAAAGTCTCCTCGTTTGCCATTGAGGAATTGTTCTTC
White_          ATGAGTGTATTGACCCAGATGCACAAAGTCTCCTCGTTTGCCATTGAGGAATTGTTCTTC
Zurich_         ATGAGTGTATTGACCCAGATGCACAAAGTCTCCTCGTTTGCCATTGAGGAATTGTTCTTC
                ************************************************************

Rutgers_        CGCAAAACTATTGGTGACATTACCATTGTTCGTCTGATTGGTGACATGTACAGCCAGCGA
Cardiff_        CGCAAAACTATTGGTGACATTACCATTGTTCGTTTGATTGGTGACATGTACAGCCAGCGA
Millan_         CGCAAAACTATTGGTGACATTACCATTGTTCGTTTGATTGGTGACATGTACAGCCAGCGA
Rentokil_       CGCAAAACTATTGGTGACATTACCATTGTTCGTTTGATTGGTGACATGTACAGCCAGCGA
Scott_          CGCAAAACTATTGGTGACATTACCATTGTTCGTTTGATTGGTGACATGTACAGCCAGCGA
White_          CGCAAAACTATTGGTGACATTACCATTGTTCGTTTGATTGGTGACATGTACAGCCAGCGA
Zurich_         CGCAAAACTATTGGTGACATTACCATTGTTCGTTTGATTGGTGACATGTACAGCCAGCGA
                ********************************** ***************************

Rutgers_        AAAATT
Cardiff_        AAAATT
Millan_         AAAATT
Rentokil_       AAAATT
Scott_          AAAATT
White_          AAAATT
Zurich_         AAAATT
                ******
```

tll intron alignment

```
Rutgers_        GTAAGTATTTCAACCAAGCAATATTCTCAACGAAGCAGGCGAATTAACAGCAAGGATGAA
Cardiff_        GTAAGTATTTCAACCAAGCAATATTCTCAACGAAGCAGGCGAATTAACAGCAAGGATGAA
Millan_         GTAAGTATTTCAACCAAGCAATATTCTCAACGAAGCAGGCGAATTAACAGCAAGGATGAA
Rentokil_       GTAAGTATTTCAACCAAGCAATATTCTCAACGAAGCAGGCGAATTAACAGCAAGGATGAA
Scott_          GTAAGTATTTCAACCAAGCTACATTCTTAACGAAGCAGGCGAATCAACAGCAAGGATGAA
White_          GTAAGTATTTCAACCAAGCAATATTCTCAACGAAGCAGGCGAATTAACAGCAAGGATGAA
Zurich_         GTAAGTATTTCAACCAAGCAATATTCTCAACGAAGCAGGCGAATTAACAGCAAGGATGAA
                ******************* * ***** *************** ***************

Rutgers_        CCAAAAGCTACTTTGCCATTTCAAAGGATTACCCAAACTT-GAAAAAGCCAAAGAAGTTT
Cardiff_        CCAAAAGCTACTTTGCCATTTCAAAGGATTACCCAAACTTTGAAAAAGCCAAAGAAGTTT
Millan_         CCAAAAGCTACTTTGCCATTTCAAAGGATTACCCAAATTTTGAAAAAGCCAAAGAAGTTT
Rentokil_       CCAAAAGCTACTTTGCCATTTCAAAGGATTACCCAAACTT-GAAAAAGCCAAAGAAGTTT
Scott_          CCAAAAGCTACTTTGCCATTTCAAAGGATTACCCAAATTTTGAAAAAGCCAAAGAAGTTT
White_          CCAAAAGCTACTTTGCCATTTCAAAGGATTACCCAAATTTTGAAAAAGCCAAAGAAGTTT
Zurich_         CCAAAAGCTACTTTGCCATTTCAAAGGATTACCCAAACTT-GAAAAAGCCAAAGAAGTTT
                ************************************ ** ******************

Rutgers_        TGCTTCCTCTCCAGCTAGGAATTCACAAAATATTTCTCGAATCTCATTTCACAATTTACT
Cardiff_        TGCTTCCTCTCCAGCTAGGAATTCACAAAATATTTCTCGAGTCTCATTTCACAATTTACT
Millan_         TGCTTCCTCTCCAGCTAGGAATTCACAAAATATTTCTCGAGTCTCATTTCACAATTTACT
Rentokil_       TGCTTCCTCTCCAGCTAGGAATTCACAAAATATTTCTCGAATCTCATTTCACAATTTACT
Scott_          TGCTTCCTCTCCAGCTAGGAATTCACAAAATATTTCTCGAGTCTCATTTCACAATTTACT
White_          TGCTTCCTCTCCAGCTAGGAATTCACAAAATATTTCTCGAGTCTCATTTCACAATTTACT
Zurich_         TGCTTCCTCTCCAGCTAGGAATTCACAAAATATTTCTCGAGTCTCATTTCACAATTTACT
                ****************************************** ******************

Rutgers_        AATTGATTTTTTCTCTCTTGCTCTAATTTTAG
Cardiff_        AATTGATTTTTTCTCTCTTGCTCTAATTTTAG
Millan_         AATTGATTTTTTCTCTCTTGCTCTAATTTTAG
Rentokil_       AATTGATTTTTTCTCTCTTGCTCTAATTTTAG
Scott_          AATTGATTTTTTCTCTCTTGCTCTAATTTTAG
White_          AATTGATTTTTTCTCTCTTGCTCTAATTTTAG
Zurich_         AATTGATTTTTTCTCTCTTGCTCTAATTTTAG
                *********************************
```

Appendix 3 – SIMPLE34 data

tll promoter region

```
*** Sequence characteristics ***   G:269  A:620  T:573  C:314
Length: 1776

*** Calculation of simplicity ***   Score mono-elements:   0
Score di-elements:   0  Score tri-elements:   1  Score tetra-
elements:   3 Size of window:  32  Simplicity was calculated from
element  33 to element 1741

*** Randomization of sequences ***   Number of generated random
sequences:  10  Randomization according to di-element frequency.

*** Simplicity ***   Simplicity test sequence 2.8403
Average simplicity random sequences: 2.4211  Ratio (test/random):
1.1731
*** Confidence limits ***   Variance simplicity random sequences:
0.0280  Standard Error: 0.0529  Confidence limit 0.95: (2.3175 -
2.5248)  Confidence limit 0.99: (2.2847 - 2.5576)

*** Significance of simplicity ***  Simplicity in sequence is
significant(confidence 0.99)
ANALYSIS OF SEQUENCE : File : Tllpromoter.sdn
PARAMETERS ARE: 3 32 1
4 32 3
ANALYSIS FROM NT 1 TO NT 1669
SIMPLICITY FACTOR 2.737
ANALYSIS OF RANDOMIZED SEQUENCE (DOUBLETS):
COMPLETE RANDOMIZATION WITH 0.342 A, 0.180 C 0.156 G AND 0.322 T
(OR U)
RANDOMIZED SIMPLICITY FACTOR 2.274 SD = 0.181
RELATIVE SIMPLICITY FACTOR 1.204
95.0% CONFIDENCE LIMITS 1.065 AND 1.385
99.0% CONFIDENCE LIMITS 1.016 AND 1.477
99.7% CONFIDENCE LIMITS 0.972 AND 1.581
HIGHEST SCORE IN RANDOMIZED RUNS: 18.00
NUMBER OF SIGNIFICANT MOTIFS: 2
THRESHOLD FOR SIGNIFICANCE: 0.999

Simplicity-rich motifs and segments:
The motifs that accumulated high simplicity scores are shown.
The number in front of the motif indicates its position
in the sequence.  Below the motifs is the correspondent
segment of the sequence where the repeats  can be found.
The motifs that accumulated high scores are marked with
asterisks.  The abundance of the different motifs is also
```

displayed at the end of the file.

```
273  TAAA
241- AATATTAAAATAAATAAATGCACACACTGTTTTAAAAACCATTTTAATTAAATTCTAAGCTAAAATCT -  308
                                      ****
372  TTTT
373  TTTT
374  TTTT
340- TGTTATTTTTATGACATCGGTCTCCCTAAGTATTTTTTTTTTTTTTCGAAATAATCTCTCTTTTGCATAT -  409
                                         ******
376  TTTT
377  TTTT
378  TTTT
344- ATTTTTATGACATCGGTCTCCCTAAGTATTTTTTTTTTTTTTCGAAATAATCTCTCTTTTGCATATTTTG -  413
                                         ******
380  TTTT
381  TTTT
382  TTTT
348- TTATGACATCGGTCTCCCTAAGTATTTTTTTTTTTTTTCGAAATAATCTCTCTTTTGCATATTTTGTAAG -  417
                                      ******

400  TTTT
368- AGTATTTTTTTTTTTTTTCGAAATAATCTCTCTTTTGCATATTTTGTAAGAATTCCAAAATAGAACTA -  435
                                      ****

642  TTTT
610- AAGATTTTGTTAATTTTATATGCGGACTCATATTTTTTTTAATTTATTTTATGCCACTCTGCCGTGTT -  677
                                         ****

644  TTTT
645  TTTT
646  TTTT
612- GATTTTGTTAATTTTATATGCGGACTCATATTTTTTTTAATTTATTTTATGCCACTCTGCCGTGTTTTGG -  681
                                      ******
Total  Motif     14    TTTT     1    TAAA
```

Significant Motifs are:

```
     271  TTTT       21    S = 1.000
     272  TTTT       21    S = 1.000
```

tll coding sequence

```
ANALYSIS OF SEQUENCE : File : Tllcoding.sdn
 PARAMETERS ARE:        3  32   1
                        4  32   3
 ANALYSIS FROM NT       1 TO NT   1326
 SIMPLICITY FACTOR                2.087

ANALYSIS OF RANDOMIZED SEQUENCE (DOUBLETS):
 COMPLETE RANDOMIZATION WITH    0.249 A, 0.290 C 0.200 G AND
0.261 T (OR U)
 RANDOMIZED SIMPLICITY FACTOR   1.990  SD =  0.156
 RELATIVE SIMPLICITY FACTOR     1.049
 95.0% CONFIDENCE LIMITS        0.930  AND  1.204
 99.0% CONFIDENCE LIMITS        0.888  AND  1.282
```

```
 99.7% CONFIDENCE LIMITS         0.850  AND  1.371
 HIGHEST SCORE IN RANDOMIZED RUNS:   16.00
 NUMBER OF SIGNIFICANT MOTIFS:     1
 THRESHOLD FOR SIGNIFICANCE:  0.999

Significant Motifs are:

1270 ATTG

1151 -  AGTATGAGATTTGTCAAAAATTGGTGACCTTTCTCAAACCCTGTCTAACCCCCTGATTGT - 1310
                       ****

Significant Motifs are:

    1270  ATTG      17   S = 1.000
```

tll 5'UTR and intron

```
ANALYSIS OF SEQUENCE : File : Tll3p.sdn
 PARAMETERS ARE:        3  32   1
                        4  32   3
 ANALYSIS FROM NT       1 TO NT    409
 SIMPLICITY FACTOR                3.787

ANALYSIS OF RANDOMIZED SEQUENCE (DOUBLETS):
 COMPLETE RANDOMIZATION WITH    0.416 A, 0.193 C 0.125 G AND
0.267 T (OR U)
 RANDOMIZED SIMPLICITY FACTOR   3.072  SD =  0.396
 RELATIVE SIMPLICITY FACTOR     1.233
 95.0% CONFIDENCE LIMITS        1.017  AND  1.564
 99.0% CONFIDENCE LIMITS        0.948  AND  1.760
 99.7% CONFIDENCE LIMITS        0.889  AND  2.009
 HIGHEST SCORE IN RANDOMIZED RUNS:   21.00
 NUMBER OF SIGNIFICANT MOTIFS:     0
 THRESHOLD FOR SIGNIFICANCE:  0.999

Significant Motifs are: 0
```

cad coding sequence

```
*** Sequence characteristics ***

 G:249  A:353  T:251  C:353

 Length: 1206

 *** Simplicity ***
```

```
 Simplicity test sequence 2.8349
 Average simplicity random sequences: 1.7177
 Ratio (test/random): 1.6504

 *** Confidence limits ***

 Variance simplicity random sequences: 0.0312
 Standard Error: 0.0559
 Confidence limit 0.95: (1.6082 - 1.8272)
 Confidence limit 0.99: (1.5736 - 1.8619)

 *** Significance of simplicity ***

Simplicity in sequence is significant(confidence 0.99)


Simplicity-rich motifs and segments:

Total  Motif

    7    ACAA
    5    AGCA
    4    CAAT
    4    CAGC
    4    AACA
    3    CATC
    3    CAAC
    1    GCAG
    1    CCAT
    1    CATT
    1    ATCA
```

cad 3utr sequence

```
*** Sequence characteristics ***

 G:269  A:586  T:571  C:353

 Length: 1779

 *** Simplicity ***

 Simplicity test sequence 4.7932
 Average simplicity random sequences: 2.0819
 Ratio (test/random): 2.3023

 *** Confidence limits ***
```

Variance simplicity random sequences: 0.0256
Standard Error: 0.0506
Confidence limit 0.95: (1.9827 - 2.1810)
Confidence limit 0.99: (1.9514 - 2.2124)

*** Significance of simplicity ***

Simplicity in sequence is significant(confidence 0.99)

Simplicity-rich motifs and segments:

Total	Motif
33	AAAA
31	TTTT
12	TAAA
7	TTTA
7	AATT
7	TTAA
7	AAAT
6	AGCA
5	CAGC
5	ATTT
5	ATTA
4	TTAT
3	GCAG
3	TTGT
3	TATT
3	TTTG
2	CACA
2	TCCC
2	CCCT
2	GTTT
2	AAAC
1	AATA

D. melanogaster Bcd and *M.* domestica *tll* promoter

	0	3.3	10	100	200	286	400	500	800	870	1,000	10,000	100,000
free DNA	1560	1814	1508	1613	1012	567	329	67					
complex 1			50	369	928	1039	729	433	41	111			
complex 2					293	521	440	306	114	253	25		
complex 3					57	194	163	233	77	141	71		
complex 4+						21	48	124	295	485	558	1534	1588
total	1560	1814	1558	1982	2290	2342	1709	1163	527	990	654	1534	1588
fraction bound		0	0.0320924	0.1861755	0.558078	0.7578992	0.8074897	0.9423903	1	1	1	1	1

	0	3.3	10	100	200	286	400	500	800	870	1,000	10,000	100,000
free DNA	1293	1452	1460	1099	1004	541	56	21		249			
complex 1			23	613	920	643	81	48		552			
complex 2				180	341	227	23	47		435			
complex 3					80	136	92	92	74	172	34		
complex 4+						29	319	331	643	123	1842	1250	1157
total	1293	1452	1483	1892	2345	1576	571	539	717	1531	1876	1250	1157
fraction bound		0	0.015509	0.4191331	0.571855	0.6567258	0.9019264	0.9610389	1	0.83736	1	1	1

	0	3.3	10	100	200	286	400	500	800	870	1,000	10,000	100,000
free DNA	1328	1453	1834	1553	1429	1096	358	262					
complex 1			20	951	995	1031	721	927					
complex 2				146	248	390	444	563	80	157	18		
complex 3					51	111	214	266	177	184	112		
complex 4+							192	106	959	1024	1508	1701	1628
total	1328	1453	1854	2650	2723	2628	1929	2124	1216	1365	1638	1701	1628
fraction bound		0	0.0107874	0.4139622	0.475211	0.5829528	0.8144116	0.8766478	1	1	1	1	1

	0	3.3	10	100	200	286	400	500	800	870	1,000	10,000	100,000
free DNA	1142	1209	1149	984	833	1074	385	162	13				
complex 1			27	397	482	105	603	422	203	148			
complex 2				20	84		350	341	267	282	5		
complex 3							75	140	117	172	92		
complex 4+							5	83	319	332	1052	1344	1065
total	1142	1209	1176	1401	1399	1179	1418	1148	919	934	1149	1344	1065
fraction bound		0	0.022959	0.297644	0.404574	0.089058	0.728490	0.858885	0.98585	1	1	1	1

M. domestica Bcd and *M.* domestica *tll* promoter

	0	3.3	10	100	200	286	400	500	800	870	1,000	10,000	100,000
free DNA	1811	1314	632	1156	472	320			92				
complex 1	1392			345	176	311	29	32	525	15			
complex 2				14	21	114	45	113	530	52			
complex 3							30	122	293	63	5		
complex 4+								188	199	296	115	213	822
total	3203	1314	632	1515	669	745	104	455	1639	426	120	213	822
fraction bound		0	0	0.2369637	0.2944693	0.5704698	1	1	0.9438682	1	1	1	1

	0	3.3	10	100	200	286	400	500	800	870	1,000	10,000	100,000
free DNA	1666	1300	1653	1181	466	388	20	389					
complex 1	1442			305	212	263	40	660					
complex 2					18	96	155	390					
complex 3							124	151	13	18			
complex 4+							42	33	120	153	191	384	112
total	3108	1300	1653	1486	696	747	381	1623	133	171	191	384	112
fraction bound		0	0	0.2052489	0.3304597	0.4805890	0.9475065	0.7603203	1	1	1	1	1

	0	3.3	10	100	200	286	400	500	800	870	1,000	10,000	100,000
free DNA	1199	1375	1431	1405	1187	855	564	511	68				
complex 1	1228			173	449	543	689	697	336	49			
complex 2					68	72	282	258	320	179			
complex 3							64	43	134	152			
complex 4+							5	13	92	409	1198	1658	440
total	2427	1375	1431	1578	1704	1470	1604	1522	950	789	1198	1658	440
fraction bound		0	0	0.1096324	0.3034037	0.4183673	0.6483790	0.6642575	0.9284210	1	1	1	1

	0	3.3	10	100	200	286	400	500	800	870	1,000	10,000	100,000
free DNA	1848	1621	1531	1524	1424	1577	1421	1211	983	576	341		
complex 1	1302			339	318	502	1065	489	1240	1082	688		
complex 2						25	263	35	464	387	312		
complex 3									51	70	120		
complex 4+											5	2139	1199
total	3150	1621	1531	1863	1742	2104	2749	1735	2738	2115	1466	2139	1199
fraction bound		0	0	0.181964	0.182548	0.250475	0.483084	0.302017	0.640978	0.727659	0.767394	1	1

M. domestica Bcd and *D.* melanogaster *tll* promoter

	0	3.3	10	100	200	286	400	500	800	870	1,000	10,000	100,000
free DNA	1227	1462	1270	1269	1431	1125	403	523	170	204	152		
complex 1	1660		15	283	371	512	232	364	5	31	18		
complex 2					58	110	166	190	5	5	96		
complex 3							91	83	57	29	108		
complex 4+							40	46	147	355	216	1573	1441
total		1462	1285	1552	1860	1747	932	1206	384	624	590	1573	1441
fraction bound		0	0.0116731	0.1823453	0.23064	0.356038	0.567596	0.566334	0.5572	0.67307	0.74237	1	1

	0	3.3	10	100	200	286	400	500	800	870	1,000	10,000	100,000
free DNA	1174	1127	1237	1011	885	607	125	164	94	90	60		
complex 1				341	457	445	146	213	18	49	20		
complex 2				50	115	251	224	260	41	70	15		
complex 3						10	241	269	65	93	110		
complex 4+							381	156	634	486	1029	1784	1403
total		1127	1237	1402	1457	1313	1117	1062	852	788	1234	1784	1403
fraction bound		0	0	0.278887	0.39258	0.5376999	0.888093	0.845574	0.889671	0.88578	0.95137	1	1

	0	3.3	10	100	200	286	400	500	800	870	1,000	10,000	100,000
free DNA	1290	1610	1746	1534	1467	1101	865	661	262	220	225		
complex 1	2012			417	514	584	466	497	135	61	35	76	
complex 2					38	90	170	267	122	85	39		
complex 3							74	169	164	110	85		
complex 4+										225	340	1246	1270
total		1610	1746	1951	2019	1775	1575	1594	683	701	724	1322	1270
fraction bound		0	0	0.213736	0.273402	0.379718	0.4507936	0.5853199	0.616398	0.686162	0.689226	1	1

	0	3.3	10	100	200	286	400	500	800	870	1,000	10,000	100,000
free DNA	1085	1424	1423	1299	1418	1318	1003	735	335	216	192		
complex 1	1762			394	490	460	521	351				110	135
complex 2					47	67	190	123	117	46	134		130
complex 3							45	48	117	44	113		
complex 4+									459	179	255	1446	1375
total		1424	1423	1693	1955	1845	1759	1257	1028	485	694	1556	1640
fraction bound		0	0	0.232722	0.274680	0.285636	0.429789	0.415274	0.674124	0.554639	0.723342	1	1

D. melanogaster Bcd and *D.* melanogaster *tll* promoter

	0	3.3	10	100	200	286	400	500	800	870	1,000	10,000	100,000
free DNA	1359	1526	1332	1227	723	628	361	232	187	85			
complex 1				321	280	270	167	196					
complex 2				13	79	197	141	115	3				
complex 3					66	87	110	119	23	8			
complex 4+							18	86	267	396	1261	1416	1176
total	1359	1526	1332	1561	1148	1182	797	748	480	489	1261	1416	1176
fraction bound		0	0	0.213965	0.3702090	0.4686971	0.5470514	0.6898395	0.610416	0.826175	1	1	1

	0	3.3	10	100	200	286	400	500	800	870	1,000	10,000	100,000
free DNA	1449	1754	1835	1574	1229	446	345	293	262	198	126		
complex 1				427	696	288	358	265	84	36	58	87	
complex 2				89	311	280	378	323	154	149	33		
complex 3					113	130	214	326	228	239	85		
complex 4+						5	53	248	759	518	373	1656	1474
total	1449	1754	1835	2090	2349	1149	1348	1455	1487	1140	675	1743	1474
fraction bound		0	0	0.246889	0.476798	0.611836	0.744065	0.798625	0.823806	0.826315	0.813333	1	1

	0	3.3	10	100	200	286	400	500	800	870	1,000	10,000	100,000
free DNA	1282	2360	2318	2164	2026	2071	888	1161	576	543	247		
complex 1	1111			497	645	820	544	856	565	477	114	218	114
complex 2					124	182	284	285	374	390	208		
complex 3							195	180	212	209	159		
complex 4+									156	86	425	2058	1749
total	2393	2360	2318	2661	2795	3073	1911	2482	1883	1705	1153	2276	1863
fraction bound		0	0	0.186771	0.275134	0.326065	0.535321	0.532232	0.694105	0.681524	0.785776	1	1

	0	3.3	10	100	200	286	400	500	800	870	1,000	10,000	100,000
free DNA	1096	1904	1993	1814	1543	1218	771	938	178	286	173		
complex 1	1645			390	553	440	291	383		10			
complex 2					61	140	169	193	46	10	33		
complex 3								20	85	65	42		
complex 4+								27	300	704	890	1335	1703
total		1904	1993	2204	2157	1798	1231	1561	609	1075	1138	1335	1703
fraction bound		0	0	0.17695	0.284654	0.322580	0.373679	0.399103	0.707717	0.733953	0.847978	1	1

Appendix 5 - Fly crosses

All transgenic lines were crossed to w118 and the marker Sb was selected against to remove the Δ2-3 immobilised P-element.
The resulting w; tll homozygous stocks were crossed to w; ScO/CyO; MKRS/TM6 to produce a balanced stocks of w; tll/CyO; MKRS/TM6 and w; ScO/CyO; tll/TM6.

To generate a stock containing the M2.2*tll-lacZ* transgene in a *bcd* mutant background the following crosses were made:
(M = male, F = female)

1 ⇒ Df(3R)LIN/TM6 × w; ScO/CyO; MKRS/TM6
⇒ select for Sb and ScO or Cy and w in males
⇒ w; ScO; Df(3R)LIN/MKRS (M) × CyO; Df(3R)LIN/MKRS (F)
⇒ select for w; ScO and Cy

⇒ w; ScO/CyO; MKRS/Df(3R)LIN

2 ⇒ w;bcdE1/TM3 × w; ScO/CyO; MKRS/TM6
⇒ remove Tb embryos, select for ScO or Cy
⇒ w; ScO; MKRS/bcdE1 × w; CyO; MKRS/bcdE1
⇒ select for ScO and Cy

⇒ w; ScO/CyO; MKRS/bcdE1

3 ⇒ w; tll/CyO; MKRS/TM6 × w; ScO/CyO; MKRS/Df(3R)LIN (1)
⇒ select against Sb and ScO

⇒ w; tll/CyO; TM6/Df(3R)LIN

4 ⇒ w; tll/CyO; MKRS/TM6 × w; ScO/CyO; MKRS/bcdE1 (2)
⇒ select against Sb and ScO
⇒ w; tll/CyO; TM6/bcdE1

5 ⇒ w; tll/CyO; TM6/Df(3R)LIN (3) × w; tll/CyO; TM6/bcdE1 (4)
⇒ remove Tb embryos

⇒ w; tll/tll; Df(3R)LIN/bcdE1

To generate a stock containing the M2.2*tll-lacZ* transgene in a *tor* mutant background the following crosses were made:

1 ⇒ Tor4/CyO × w; ScO/CyO; MKRS/TM6
⇒ select for Sb and against ScO

⇒ Tor/CyO; MKRS

2 ⇒ Tor4/CyO; MKRS (1) × w; ScO/CyO; tll/TM6
⇒ remove Tb embryos and select for Sb and against ScO

⇒ Tor4/CyO; tll/MKRS

3 ⇒ self cross of Tor4/CyO; tll/MKRS (2)
⇒ select against Cy and Sb

⇒ Tor4/Tor4; tll/tll

Appendix 6 - *cad* sequence

Length: 3371
Translation start: 388
Inton position: 1152
Stop codon: 1590

```
   1  TCTAGATCTA GACTCAAATC CCCCCAAAGA AGAATATCCA ATAATAACAA
  51  AAAATAAACG AAAAAAGCAT TAAAAGGTTT TGCAACAAAA CCAAGACAAC
 101  AACGCGTGCA AGACAAGAAA ATaAAAAAAA aGACGAAAGG GAGAAATaAA
 151  AAACAATTGA GTGtTTAAGA AATAGAAATT ATTGCAAATG AAATGaaAAA
 201  AAAATGTAAA ACATTGTAAA TAAtAaAcAA AAAACAAACg ACAACTacAA
 251  AAATTAtAAA ATAATACaAA CACCAAGTTT cTTcGCCcCC AATtATCCCA
 301  GTGACAATcT AtTTcTAACC TCAAAgTTTT CTTCTTCTTC TTCGacgacg
 351  agtAAAgACA AgCCGCGTGA CCCACACCac aACCAtCATG GTcTCaTTCT
 401  ACAACACCCT ACCaTATACG CAAAAgCACA GCGCCAATTT GGCCTATTcC
 451  GCCGGACAAC CCTGGCAATG GACGGCCAAT TATCATCACA CGCCaCCCAA
 501  TCAtCAATAT CTGAGCGACA TGGACTCAAC ACATGCCGcT GcGGCCCATC
 551  ATCAAATGTA CTATAATCCT CATGCCATGT ATCATTcGGC CACAAATGCt
 601  GCTGCGGCCg CCGCCTcAGG TTGGCATTCC CccTcGTcGG cGGAGAaTTT
 651  CTCACAAAAT TCCCAATTGT TGAGCCAACA ACATCAACAG CTCCTAAATG
 701  GTACCGTcGT TGGtGGcGGT GCAACACCAT CATCATCGTC GGCCAGTGCC
 751  AGTAGTACAA CATCGGCAGG TCCAGCGTCT GGCAGTACGA CACAATTGAA
 801  CGAGACCGTT AGCAGtATTG GcGATGTCCA GCATCCGCAG CAGCAaCAAC
 851  AACAACAGCA GCAGGCCCAG CAACAGGCTC ATCATCACAT CACCGAaGGa
 901  TTGCCATCGC CgCCCATtAC cGTtAGTGGC AGTGAaATAT CCAGtCCGGG
 951  AGCCCCAGCC TCATCaTCAT CGCCGAATCA CATTGCCCAT CATTTGAaTA
1001  ATAACCATAG TCCATCGACa GCCAATAAtA aCAACAaCAa TACCATAAAT
1051  CACAaCAaCA aCAATCGTTC GTCACCGGTG AAATCGCATC AGTACTACGA
1101  TTGGATGAAG AAACCAACAT ATCCGGCCCA GCCAGCACCt GGTAAAACCC
```

1151 GCAGGrwwAA ttaCsgcktG kTkwmmmCsr myTTymassk TTkGrAwTkG

1201 rAAAArrAta mtgtacatca cgctacatya cCatacgacg caAAAcCgaA

1251 tTgGsCCaaa cgctatcgct gtcgGagcgc CagGtgaAga tatgGtTTCa

1301 AaAtCgtCgc gcCaAGGAAC GCAAACAAAA CAAGAAAGTC AGTGAGCCCA

1351 GCATTGGCGG TGTCCAGCAT CCCGACTATG CCAATTTGAT GGATACcAAA

1401 CCGAAACTGG AACCCGGTAT ACATTTGCAA CATTcCCTGC ATTCGATGGC

1451 TGCCATGGGT ATGCCAGCAA TGCGTTTACA TcCACATTTG CATGGGCATC

1501 ATCATTTGGc TGTGAGTGCT GCcCATTCAC ATCAATTGCA TCAATCGCCG

1551 CATGCCCAGA TATCAGCGGC TGTGGGTAGC CTATCGATGT GACAACCTTA

1601 TGCGACAAGC TTAGATGGCG GCAGCCTAGT TTGcCCAAAT AGCAATAATC

1651 ATAATTTTAT AAATAATCAG AGCAGCAGCA GCAaCAACAA CagcAACAAT

1701 AGCAGCGGCA GCGGCAGCAT CCACCATGGA GGTGATGAGG AACAAAGTAT

1751 TTTGCAtGAT CTAgCAgCTG TGGCGTTGGG CGCCGATGAT GTTGAAAGTG

1801 CATTGGCCGT TGTGGCGGCA GCTGCTGCTG TCACTGGCGC CGACGGCAAT

1851 GACAATGTTA ATGGTGACAG TACCGTCGGC GGCGGcCATT aTATCAGCAT

1901 CATGTCTAGC ATGCAATAAT CCAACCTGTC CTAGTGAATT ATTATTACCA

1951 CTTGGCCCAT CGCACTGCCA CACAGCttag ataTaTagca cAggcGAGAA

2001 CCACACCTGC ACCCAGCCAC AGCCACAGTG ACGATTGGTC AATATCCTAA

2051 CAGGGTTATA TAATATTTAA TGTCTAGTTT TAGTGTTTAT TTTTTGTTAG

2101 TTTGTAAGTT CCCCAAAGAG TTGGGTGGCA CCCCGATATC ACCTTGTGGC

2151 CGATAATGCA AGTGGATGCG TGTGTGTGAG AGAGTGAATG TAATCATTTT

2201 GTCTGTATGT GTAAATAGTT GTTATTTTAG TATGTCAACT GCAAGTTGAA

2251 TATTGCACGT GTTGTCTTTT GTGATTGtAT ATTCCAACTG TTGTTGAGAC

2301 GTACTAATTT ATCGTCCCCC TccCCCCaTC tcCCtCcTTT TCCCTgAAAA

2351 CCCCtTTgTG TGGtTACATT CCTTCTCCTG CaCACGCCTT GCATTTTTGg

2401 CAaTGTGTTA TGTTTTtGGC CATCTCTTCT CCtCaCCaCC CCCAATCCCC

2451 CTTTATATTA AGTAAAAACC ATTTATGTGT TTTTTTtATt ATTATTGtTT

2501 TTTTGtTTTG AGTTTTTTTA TTTTTAGTTT TTTTTttGAG TTATTTTTAT

2551 TTAATTTTCC AAtTTTTTTT GTCACCATAA TATTGAACAA GAAAAAACAA

2601 ATAATTAAGC TAATATTAAG TTTACTACTA CTACTATGCC TTACAACGCA

2651 CACACACACA CCCATACATA GAGAGAGAGA GAAATATCAA ACATTTGTAG

2701 CCGTCCTTTA AAAAAAAAAC AAAAaCAAAA GAGATCaCAA AAAAaCCaCA

2751 ACTCGTAAGA TTAAGTTAAA AAAAaAaAAc TaaaTAATAT TAAATTAAAA

2801 AGGGGaGTgc AAAgGGTTAT AAGgCGACGA AATCGTTTAC tACTAcCTCC

2851 CATAAATGGT TAGTTTTCGC AATTCCAATT GGTTTTCTTT TAAcTATTCC

2901 AATTGGGAAA tTCAAATTtC AAaTcGAtAg cTTaCaaCCC cAAgCgCCCC

2951 CTcTTTGTAT tTATAAGTGA GAAaAAAAaa AAAcAaacaa cATCAATTAA

3001 CGAATTTTTG TTCAATACTG TGCAATTATC TATAACTACA ACACACAtct

3051 ACaCtCtatt tCtCCCCCtg CCCCACTTAC TaATCTTATG TAAATAtaaT

3101 ccccccacGA CCCCTCCCAG AACACAAATG AATTTCTAGT TTCtATTTCt

3151 AtTATTtAtG tGGtCATTCt TTtTaTTTTT TTTTTaAATA TAGCCTATAT

3201 TTAATTAACG TTTATTACTT ATTAAATAAT AAGaAaCAAA TTAGCaACAA

3251 CAaCAaCAAA AATATTTCTC TAATTTAGTT AAAATTAAaT AAAAaTtAaA

3301 ATTAAAATTA AAATTATAAT AAAAAAGCTT AAAGACAAAA TAAATCtaCT

3351 ACTTATaaTA AAAAAAAAAA A

References

Ackers, G. K., Shea, M. A., and Smith, F. R. (1983). Free-energy coupling within macromolecules - the chemical work of ligand binding at the individual sites in co-operative systems. *Journal of Molecular Biology* **170**, 223-242.

Ades, S. E. and Sauer, R. T. (1994). Differential DNA-binding specificity of the engrailed homeodomain: the role of residue 50. *Biochemistry* **33**, 9187-9194.

Ades, S. E. and Sauer, R. T. (1995). Specificity of minor-groove and major-groove interactions in a homeodomain-DNA complex. *Biochemistry* **34**, 14601-14608.

Akam M. (1994). Insect development. Is pairing the rule? *Nature* **367**:410-1.

Akashi H. (2001). Gene expression and molecular evolution. *Curr Opin Genet Dev.* **11**:660-6.

Amores A, Force A, Yan YL, Joly L, Amemiya C, Fritz A, Ho RK, Langeland J, Prince V, Wang YL, Westerfield M, Ekker M, Postlethwait JH. (1998). Zebrafish *hox* clusters and vertebrate genome evolution. *Science* **282**:1711-4.

Aquadro CF, Bauer DuMont V, Reed FA. (2001) Genome-wide variation in the human and fruitfly: a comparison. *Curr Opin Genet Dev.* **11**:627-34.

Arnone, M. I. and Davidson, E. H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. *Development* **124**, 1851-1864.

Arnosti, D. N., Barolo, S., Levine, M., and Small, S. (1996). The *eve* stripe 2 enhancer employs multiple modes of transcriptional synergy. *Development* **122**, 205-214.

Arthur, W (2002). The emerging conceptual framework of evolutionary developmental biology. *Nature* **415**:757-64

Baines JF, Chen Y, Das A, Stephan W. (2002). DNA sequence variation at a duplicated gene: excess of replacement polymorphism and extensive haplotype structure in the *Drosophila melanogaster bicoid* region. *Mol Biol Evol.* **19**:989-98.

Bellaïche, Y., Bandyopadhyay, R., Desplan, C., and Dostatni, N. (1996). Neither the homeodomain nor the activation domain of Bicoid is specifically required for its down-regulation by the Torso receptor tyrosine kinase cascade. *Development* **122**, 3499-3508.

Belting, H., Shashikant, C. S. and Ruddle, F. H. (1998). Modification of expression and *cis*-regulation of *Hoxc8* in the evolution of diverged axial morphology. *PNAS USA* **95**, 2355-2360.

Bergman CM, Kreitman M. (2001) Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res.* **11**:1335-45

Berleth, T., Burri, M., Thoma, G., Bopp, D., Richstein, S., Frigerio, G., Noll, M. and Nüsslein-Volhard, C. (1988). The role of localization of *bicoid* RNA in organizing the anterior pattern of the *Drosophila* embryo. *EMBO J* **7**, 1749-1756.

Beverley, S. M. and Wilson, A. C. (1984). Molecular evolution in *Drosophila* and the higher Diptera II. A time scale for fly evolution. *J Mol Evol* **21**, 1-13.

Bonneton, F., Theodore, L., Silar, P., Maroni, G. and Wegnez, M. (1996). Response of *Drosophila* metallothionein promoters to metallic, heat shock and oxidative stresses. *FEBS Lett* **380**, 33-38.

Bonneton, F., Shaw, P. J., Fazakerley, C., Shi, M. and Dover, G. A. (1997). Comparison of *bicoid*-dependent regulation of *hunchback* between *Musca domestica* and *Drosophila melanogaster*. *Mech Dev* **66**, 143-156.

Breiling A, Turner BM, Bianchi ME, Orlando V. (2001) General transcription factors bind promoters repressed by Polycomb group proteins. *Nature.* **412**:651-5

Brown, S., Fellers, J., Shippy, T., Denell, R., Stauber, M. and Schmidt-Ott, U. (2001). A strategy for mapping *bicoid* on the phylogenetic tree. *Curr Biol* **11**, R43-44.

Burke TW, Kadonaga JT. (1996) Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. *Genes Dev.* **10**:711-24.

Burke TW, Kadonaga JT. (1997). The downstream core promoter element, DPE, is conserved from *Drosophila* to humans and is recognized by TAFII60 of *Drosophila*. *Genes Dev.* **11**:3020-31.

Burz, D. S., Rivera-Pomar, R., Jäckle, H. and Hanes, S. D. (1998). Cooperative DNA-binding by Bicoid provides a mechanism for threshold- dependent gene activation in the *Drosophila* embryo. *EMBO J* **17**, 5998-6009.

Burz, D. S. and Hanes, S. D. (2001). Isolation of mutations that disrupt cooperative DNA binding by the *Drosophila bicoid* protein. *J Mol Biol* **305**, 219-230.

Carroll, S. B., Grenier, J. K. and Weatherbee, S. D. (2001). *From DNA to Diversity, Molecular Genetics and the Evolution of Animal Design*. Malden: Blackwell Science.

Cavener DR. (1987) Comparison of the consensus sequence flanking translational start sites in *Drosophila* and vertebrates. *Nucleic Acids Res.* **15**:1353-61.

Chen YJ, Chiang CS, Weng LC, Lengyel JA, Liaw GJ. (2002) Tramtrack69 is required for the early repression of *tailless* expression. *Mech Dev.* **116**:75-83.

Cherbas, L. and Cherbas, P. (1993). The arthropod initiator: the capsite consensus plays an important role in transcription. *Insect Biochem Mol Biol* **23**, 81-90.

Church, G. M. and Gilbert, W. (1984). Genomic sequencing. *PNAS USA* **81**, 1991-1995.

Crain WR, Davidson EH, Britten RJ. (1976) Contrasting patterns of DNA sequence arrangement in *Apis mellifera* (honeybee) and *Musca domestica* (housefly). *Chromosoma.* **59**:1-12.

Dave, V., Zhao, C., Yang, F., Tung, C. S. and Ma, J. (2000). Reprogrammable recognition codes in *bicoid* homeodomain-DNA interaction. *Mol Cell Biol* **20**, 7673-84.

Davidson, E. H. (2001). *Genomic regulatory systems : development and evolution*. San Diego: Academic Press.

Davis AP, Capecchi MR. (1996) A mutational analysis of the 5' *HoxD* genes: dissection of genetic interactions during limb development in the mouse. *Development.* **122**:1175-85.

Dearden, P. and Akam, M. (1999). Developmental evolution: Axial patterning in insects. *Curr Biol* **9**, R591-594.

Dermitzakis ET, Clark AG. (2002) Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol Biol Evol.* **19**:1114-21.

Devon, R. S., Porteous, D. J. and Brookes, A. J. (1995). Splinkerettes improved vectorettes for greater efficiency in PCR walking. *Nucleic Acids Res* **23**, 1644-1645.

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

Dover, G. A. and Flavell, R. B. (1984). Molecular coevolution: DNA divergence and the maintenance of function. *Cell* **38**, 622-623.

Dover, G. A. (1993). Evolution of genetic redundancy for advanced players. *Curr Opin Genet Dev* **3**, 902-910.

Dover, G. (2000). How genomic and developmental dynamics affect evolutionary processes. *Bioessays* **22**, 1153-1159.

Driever, W. and Nüsslein-Volhard, C. (1988a). A gradient of *bicoid* protein in *Drosophila* embryos. *Cell* **54**, 83-93.

Driever, W. and Nüsslein-Volhard, C. (1988b). The *bicoid* protein determines position in the *Drosophila* embryo in a concentration-dependent manner. *Cell* **54**, 95-104.

Driever, W., Thoma, G. and Nüsslein-Volhard, C. (1989a). Determination of spatial domains of zygotic gene expression in the *Drosophila* embryo by the affinity of binding sites for the *bicoid* morphogen. *Nature* **340**, 363-367.

Driever, W., Ma, J., Nüsslein-Volhard, C. and Ptashne, M. (1989b). Rescue of *bicoid* mutant *Drosophila* embryos by *bicoid* fusion proteins containing heterologous activating sequences. *Nature* **342**, 149-154.

Driever, W. and Nüsslein-Volhard, C. (1989). The *bicoid* protein is a positive regulator of *hunchback* transcription in the early *Drosophila* embryo. *Nature* **337**, 138-143.

Dubnau, J. and Struhl, G. (1996). RNA recognition and translational regulation by a homeodomain protein. *Nature* **379**, 694-699.

Dubnicoff T, Valentine SA, Chen G, Shi T, Lengyel JA, Paroush Z, Courey AJ. (1997) Conversion of dorsal from an activator to a repressor by the global corepressor Groucho. *Genes Dev.* **11**:2952-7.

Emili, A., Greenblatt, J. and Ingles, C. J. (1994). Species-specific interaction of the glutamine-rich activation domains of SP1 with the TATA box-binding protein. *Mol Cell Biol* **14**, 1582-1593.

Espinas, M. L., Canudas, S., Fanti, L., Pimpinelli, S., Casanova, J. and Azorin, F. (2000). The GAGA factor of *Drosophila* interacts with SAP18, a Sin3-associated polypeptide. *EMBO Reports* **1**, 253-259.

Falciani F, Hausdorf B, Schröder R, Akam M, Tautz D, Denell R, Brown S. (1996) Class 3 *Hox* genes in insects and the origin of *zen*. *Proc Natl Acad Sci U S A.* **93**:8479-84.

Feinberg, A. P. and Vogelstein, B. (1984). "A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity". Addendum. *Anal Biochem* **137**, 266-267.

Ferrandon, D., Elphick, L., Nüsslein-Volhard, C., and St. Johnston, D. (1994). STAUFEN protein asociates with the 3'UTR of *bicoid* messenger-RNA to form particles that move in a microtubule-dependent manner. *Cell* **79**, 1221-1232.

Finkelstein, R. and Perrimon, N. (1990). The *orthodenticle* gene is regulated by *bicoid* and *torso* and specifies *Drosophila* head development. *Nature* **346**, 485-488.

Finkelstein, R., Smouse, D., Capaci, T. M., Spradling, A. C. and Perrimon, N. (1990). The *orthodenticle* gene encodes a novel homeodomain protein involved in the development of the *Drosophila* nervous system and ocellar visual structures. *Genes Dev* **4**, 1516-1527.

Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*. **151**:1531-45. Review.

Frohnhöfer, H. G. and Nüsslein-Volhard, C. (1986). Organization of anterior pattern in the *Drosophila* embryo by the maternal gene *bicoid*. *Nature* **324**, 120-125.

Galant R, Carroll SB. (2002) Evolution of a transcriptional repression domain in an insect Hox protein. *Nature*. **415**:910-3.

Galas, D. J. and Schmitz, A. (1978). DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* **5**, 3157-3170.

Gao Q, Wang Y, Finkelstein R. (1996) Orthodenticle regulation during embryonic head development in Drosophila. *Mech Dev*. **56**:3-15

Gao, Q. and Finkelstein, R. (1998). Targeting gene expression to the head: the *Drosophila orthodenticle* gene is a direct target of the Bicoid morphogen. *Development* **125**, 4185-4193.

Gehring, W. J., Qian, Y. Q., Billeter, M., Furukubo-Tokunaga, K., Schier, A. F., Resendez-Perez, D., Affolter, M., Otting, G. and Wuthrich, K. (1994). Homeodomain-DNA recognition. *Cell* **78**, 211-223.

Gibson, G. (1996). Epistasis and pleiotropy as natural properties of transcriptional regulation. *Theor Popul Biol* **49**, 58-89.

Gibson G, van Helden S. (1997) Is function of the Drosophila homeotic gene Ultrabithorax canalized? *Genetics.* **147**:1155-68.

Gibson G, Wemple M, van Helden S. (1999) Potential variance affecting homeotic Ultrabithorax and Antennapedia phenotypes in Drosophila melanogaster. *Genetics.* **151**:1081-91.

Gray S, Levine M. (1996) Transcriptional repression in development. *Curr Opin Cell Biol.* **8**:358-64.

Greer JM, Puetz J, Thomas KR, Capecchi MR. (2000) Maintenance of functional equivalence during paralogous Hox gene evolution. *Nature.* **403**:661-5

Grossniklaus U, Cadigan KM, Gehring WJ. (1994) Three maternal coordinate systems cooperate in the patterning of the Drosophila head. *Development.* **120**:3155-71.

Hamilton, B. A., Palazzolo, M. J., Chang, J. H., VijayRaghavan, K., Mayeda, C. A., Whitney, M. A. and Meyerowitz, E. M. (1991). Large scale screen for transposon insertions into cloned genes. *PNAS USA* **88**, 2731-2735.

Hancock, J. M. and Dover, G. A. (1990). 'Compensatory slippage' in the evolution of ribosomal RNA genes. *Nucleic Acids Res* **18**, 5949-5954.

Hancock, J. M. and Armstrong, J. S. (1994). SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences. *Comput Appl Biosci* **10**, 67-70.

Hancock, J. M. (1995). The contribution of slippage-like processes to genome evolution. *J Mol Evol* **41**, 1038-1047.

Hancock, J. M. (1996). Simple sequences and the expanding genome. *Bioessays* **18**, 421-425.

Hancock, J. M., Shaw, P. J., Bonneton, F. and Dover, G. A. (1999). High sequence turnover in the regulatory regions of the developmental gene *hunchback* in insects. *Mol Biol Evol* **16**, 253-265.

Hancock, J. M. and Vogler, A. P. (2000). How slippage-derived sequences are incorporated into rRNA variable- region secondary structure: implications for phylogeny reconstruction. *Mol Phylogenet Evol* **14**, 366-374.

Hanes, S. D. and Brent, R. (1989). DNA specificity of the bicoid activator protein is determined by homeodomain recognition helix residue 9. *Cell* **57**, 1275-1283.

Hanes, S. D. and Brent, R. (1991). A genetic model for interaction of the homeodomain recognition helix with DNA. *Science* **251**, 426-430.

Hanes, S. D., Riddihough, G., Ish-Horowicz, D. and Brent, R. (1994). Specific DNA recognition and intersite spacing are critical for action of the *bicoid* morphogen. *Mol Cell Biol* **14**, 3364-3375.

Harr, B., Zangerl, B. and Schlötterer, C. (2000). Removal of microsatellite intruptions by DNA replication slippage: phylogenetic evidence from *Drosophila*. *Mol Biol Evol* **17**, 1001-1009.

Heaphy S, Dingwall C, Ernberg I, Gait MJ, Green SM, Karn J, Lowe AD, Singh M, Skinner MA. (1990) HIV-1 regulator of virion expression (Rev) protein binds to an RNA stem-loop structure located within the Rev response element region. *Cell.* **60**:685-93.

Hediger, M., Niessen, M., Wimmer, E. A., Dübendorfer, A. and Bopp, D. (2001). Genetic transformation of the housefly *Musca domestica* with the lepidopteran derived transposon *piggyBac*. *Insect Molecular Biology* **10**, 113-119.

Hoch, M., Seifert, E. and Jäckle, H. (1991). Gene expression mediated by *cis*-acting sequences of the *Krüppel* gene in response to the *Drosophila* morphogens *bicoid* and *hunchback*. *EMBO J* **10**, 2267-2278.

Holland, P. W. H., Garcia-Fernandez, J., Williams, N. A. and Sidow, A. (1994). Gene duplications and the origins of vertebrate development. *Development* supplement, 125-133.

Hollemann T, Bellefroid E, Pieler T. (1998) The Xenopus homologue of the Drosophila gene tailless has a function in early eye development. *Development.* **125**:2425-32

Houchmandzadeh B, Wieschaus E, Leibler S. (2002) Establishment of developmental precision and proportions in the early Drosophila embryo. *Nature.* **415**:798-802.

Hutson SF, Bownes M. (2003) The regulation of *yp3* expression in the *Drosophila melanogaster* fat body. *Dev Genes Evol.* **213**:1-8.

Ip YT, Park RE, Kosman D, Yazdanbakhsh K, Levine M. (1992) *dorsal-twist* interactions establish *snail* expression in the presumptive mesoderm of the *Drosophila* embryo. *Genes Dev.* **6**:1518-30

Irish, V., Lehmann, R. and Akam, M. (1989). The *Drosophila* posterior-group gene *nanos* functions by repressing *hunchback* activity. *Nature* **338**, 646-648.

Jackson A, Panayiotidis P, Foroni L. (1998) The human homologue of the *Drosophila tailless* gene (TLX): characterization and mapping to a region of common deletion in human lymphoid leukemia on chromosome 6q21. *Genomics.* **50**:34-43.

Janody, F., Sturny, R., Catala, F., Desplan, C. and Dostatni, N. (2000). Phosphorylation of *bicoid* on MAP-kinase sites: contribution to its interaction with the *torso* pathway. *Development* **127**, 279-289.

Janody, F., Sturny, R., Schaeffer, V., Azou, Y. and Dostatni, N. (2001). Two distinct domains of Bicoid mediate its transcriptional downregulation by the Torso pathway. *Development* **128**, 2281-2290.

Jeffreys, A. J., Neumann, R. and Wilson, V. (1990). Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell* **60**, 473-485.

Jenkins DL, Ortori CA, Brookfield JF. (1995) A test for adaptive change in DNA sequences controlling transcription. *Proc R Soc Lond B Biol Sci.* **261**:203-7.

Jensen MA, Charlesworth B and Kreitman M. (2002) Patterns of genetic variation at a chromosome 4 locus of *Drosophila melanogaster* and *D. simulans. Genetics.* **160**:493-507.

Jimenez G, Guichet A, Ephrussi A, Casanova J. (2000) Relief of gene repression by torso RTK signaling: role of *capicua* in *Drosophila* terminal and dorsoventral patterning. *Genes Dev.* **14**:224-31

Keys DN, Lewis DL, Selegue JE, Pearson BJ, Goodrich LV, Johnson RL, Gates J, Scott MP, Carroll SB. (1999) Recruitment of a *hedgehog* regulatory circuit in butterfly eyespot evolution. *Science.* **283**:532-4.

Kim J. (2001) Macro-evolution of the *hairy* enhancer in *Drosophila* species. *J Exp Zool.* **291**:175-85.

Kimura, M. (1983). *The neutral theory of molecular evolution.* Cambridge ; New York: Cambridge University Press.

Kobayashi M, Yu RT, Yasuda K, Umesono K. (2000) Cell-type-specific regulation of the retinoic acid receptor mediated by the orphan nuclear receptor TLX. *Mol Cell Biol.* **20**:8731-9.

Kreitman M. (2000) Methods to detect selection in populations with applications to the human. *Annu Rev Genomics Hum Genet.* **1**:539-59. Review.

Krumlauf R. (1994) *Hox* genes in vertebrate development. *Cell.* **78**:191-201. Review.

Lall S, Ludwig MZ, Patel NH.(2003) Nanos Plays a Conserved Role in Axial Patterning outside of the *Diptera. Curr Biol.* **13**:224-9.

Lawrence, P. A. (1992). The making of a fly : the genetics of animal design (Oxford [England] ; Cambridge, Mass., USA, Blackwell Science).

Lee TI, Young RA. (2000) Transcription of eukaryotic protein-coding genes. *Annu Rev Genet.* 34:77-137. Review.

Levinson, G. and Gutman, G. A. (1987). Slipped-strand mispairing - a major mechanism for DNA-sequence evolution. *Mol Biol Evol* **4**, 203-221.

Lewis EB. (1978) A gene complex controlling segmentation in Drosophila. *Nature.* **276**:565-70.

Liaw, G.-J., and Lengyel, J. A. (1992). Control of *tailless* expression by bicoid, dorsal and synergistically interacting terminal system regulatory elements. *Mechanisms of Development* **40**, 47-61.

Liaw, G. J., Steingrimsson, E., Pignoni, F., Courey, A. J., and Lengyel, J. A. (1993). Characterisation of downstream elements in a RAF-1 pathway. *Proc Natl Acad Sci USA* **90**, 858-862.

Liaw, G.-J., Rudolph, K. M., Huang, J.-D., Dubnicoff, T., Courey, A. J., and Lengyel, J. A. (1995). The torso response element binds GAGA and NTF-1/Elf-1 and regulates *tailless* by relief of repression. *Genes and Development* **9**, 3163-3176.

Lin, K. C. and Shiuan, D. (1995). A simple method for DNAseI footprint analysis. J Biochem *Biophys Methds* **30**, 85-89.

Liu S, Jack J.(1992) Regulatory interactions and role in cell type specification of the Malpighian tubules by the *cut, Kruppel,* and *caudal* genes of *Drosophila. Dev Biol.* **150**:133-43.

Ludwig, M. Z. and Kreitman, M. (1995). Evolutionary dynamics of the enhancer region of *even-skipped* in *Drosophila. Mol Biol Evol* **12**, 1002-1011.

Ludwig, M. Z., Patel, N. H. and Kreitman, M. (1998). Functional analysis of *eve* stripe 2 enhancer evolution in *Drosophila*: rules governing conservation and change. *Development* 125, 949-958.

Ludwig, M. Z., Bergman, C., Patel, N. H. and Kreitman, M. (2000). Evidence for stabilizing selection in a eukaryotic enhancer element. *Nature* **403**, 564-567.

Ludwig MZ. (2002) Functional evolution of noncoding DNA. *Curr Opin Genet Dev.* **12**:634-9.

Luk, S. K. S., Kilpatrick, M., Kerr, K. and Macdonald, P. M. (1994). Components acting in localisation of *bicoid* messenger RNA are conserved among *Drosophila* species. *Genetics* **137**, 521-530.

Lukowitz, W., Schröder, C., Glaser, G., Hülskamp, M. and Tautz, D. (1994). Regulatory and coding regions of the segmentation gene *hunchback* are functionally conserved between *Drosophila virilis* and *Drosophila melanogaster*. *Mech Dev* **45**, 105-115.

Lynch M, Force A. (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics.* **154**:459-73

Ma, X., Yuan, D., Diepold, K., Scarborough, T. and Ma, J. (1996). The *Drosophila* morphogenetic protein Bicoid binds DNA cooperatively. *Development* **122**, 1195-1206.

Ma, X., Yuan, D., Scarborough, T. and Ma, J. (1999). Contributions to gene activation by multiple functions of Bicoid. *Biochem J* **338**, 447-455.

Macdonald, P. M., and Struhl, G. (1986). A molecular gradient in early Drosophila embryos and its role in specifying the body pattern. *Nature* **324**, 537-545.

Macdonald, P. M. and Struhl, G. (1988). *cis*-acting sequences responsible for anterior localization of *bicoid* mRNA in *Drosophila* embryos. *Nature* **336**, 595-598.

Macdonald PM, Kerr K, Smith JL, Leask A. (1993) RNA regulatory element BLE1 directs the early steps of *bicoid* mRNA localization. *Development.* **118**:1233-43.

Mader S, Kumar V, de Verneuil H, Chambon P. (1989) Three amino acids of the oestrogen receptor are essential to its ability to distinguish an oestrogen from a glucocorticoid-responsive element. *Nature.* **338**:271-4

Mao, C., Carlson, N. G. and Little, J. W. (1994). Cooperative DNA-protein interactions Effects of changing the spacing between adjacent binding sites. *J Mol Biol* **235**, 532-544.

Marom K, Shapira E, Fainsod A. (1997) The chicken *caudal* genes establish an anterior-posterior gradient by partially overlapping temporal and spatial patterns of expression. *Mech Dev.* **64**:41-52.

Martin CH, Mayeda CA, Davis CA, Ericsson CL, Knafels JD, Mathog DR, Celniker SE, Lewis EB, Palazzolo MJ. (1995) Complete sequence of the bithorax complex of *Drosophila. Proc Natl Acad Sci U S A.* **92**:8398-402

McGregor, A. P., Shaw, P. J., Hancock, J. M., Bopp, D., Hediger, M., Wratten, N. S. and Dover, G. A. (2001). Rapid restructuring of bicoid-dependent *hunchback* promoters within and between Dipteran species: implications for molecular co-evolution. *Evol Dev* **3**, 397-407.

McGregor, A. P. (2002). PhD Thesis. *The evolution Bicoid regulated genes in insects.* In Department of Genetics. Leicester: University of Leicester.

Mlodzik, M., Fjose A and Gehring, W. J. (1985). Isolation of *caudal*, a *Drosophila* homeobox-containing gene with maternal expression, whose transcripts form a concentration gradient at the pre-blastoderm stage. *EMBOJ.* **4**:2961-2969.

Mlodzik, M., and Gehring, W. J. (1987). Expression of the *caudal* gene in the germline of *Drosophila* : formation of an RNA and protein gradient during early embryogenesis. *Cell* **48**, 465-478.

Monaghan AP, Grau E, Bock D, Schutz G. (1995) The mouse homolog of the orphan nuclear receptor *tailless* is expressed in the developing forebrain. *Development.* **121**:839-53.

Moriyama EN, Powell JR. (1996) Intraspecific nuclear DNA variation in *Drosophila. Mol Biol Evol.* **13**:261-77.

Murata, Y. and Wharton, R. P. (1995). Binding of pumilio to maternal *hunchback* mRNA is required for posterior patterning in *Drosophila* embryos. *Cell* **80**, 747-756.

Nagy, L. M. (1994). Insect segmentation. A glance posterior. *Curr Biol* **4**, 811-814.

Namba, R., Pazdera, T. M., Cerrone, R. L. and Minden, J. S. (1997). *Drosophila* embryonic pattern repair: how embryos respond to *bicoid* dosage alteration. *Development* **124**, 1393-1403.

Niessing, D., Dostatni, N., Jackle, H. and Rivera-Pomar, R. (1999). Sequence interval within the PEST motif of Bicoid is important for translational repression of *caudal* mRNA in the anterior region of the *Drosophila* embryo. *EMBO J* **18**, 1966-1973.

Niessing, D., Driever, W., Sprenger, F., Taubert, H., Jackle, H. and Rivera-Pomar, R. (2000) *Molecular Cell*, **5**, 395-401.

Niessing D, Blanke S, Jackle H. (2002) Bicoid associates with the 5'-cap-bound complex of *caudal* mRNA and represses translation. *Genes Dev.* **16**:2576-82.

Nusslein-Volhard C, Frohnhofer HG, Lehmann R. (1987) Determination of anteroposterior polarity in Drosophila. *Science.*; **238**:1675-81

O'Brochta, D. A., Warren, W. D., Saville, K. J. and Atkinson, P. W. (1996). Hermes, a functional non-drosophilid insect gene vector from *Musca domestica*. *Genetics* **142**, 907-914.

Ohta, T. and Dover, G. A. (1984). The cohesive population genetics of molcular drive. *Genetics* **108**, 501-521.

Oro AE, Umesono K, Evans RM. (1989) Steroid hormone receptor homologs in development. *Development.* **107** Suppl:133-40. Review

Padegimas, L. S. and Reichert, N. A. (1998). Adaptor ligation-based polymerase chain reaction-mediated walking. *Anal Biochem* **260**, 149-153.

Pan D, Courey AJ. (1992) The same dorsal binding site mediates both activation and repression in a context-dependent manner. *EMBO J.* **11**:1837-42.

Paroush Z, Wainwright SM, Ish-Horowicz D. (1997) Torso signalling regulates terminal patterning in Drosophila by antagonising Groucho-mediated repression. *Development.* **124**:3827-34.

Perrimon, N. (1993). The Torso receptor protein-tyrosine kinase signaling pathway: An endless story. *Cell* **74**, 219-222.

Piccin A, Couchman M, Clayton JD, Chalmers D, Costa R, Kyriacou CP. (2000) The clock gene *period* of the housefly, *Musca domestica*, rescues behavioral rhythmicity in *Drosophila melanogaster*. Evidence for intermolecular coevolution? *Genetics.* **154**:747-58

Pignoni, F., Balderelli, R. M., Steingrimsson, E., Diaz, R. J., Patapoutian, A., Merriam, J. R., and Lengyel, J. A. (1990). The *Drosophila* gene *tailless* is expressed at the embryonic termini and is a member of the steroid receptor superfamily. *Cell* **62**, 151-163.

Pignoni, F., Steingrimsson, E. and Lengyel, J. A. (1992). *bicoid* and the terminal system activate *tailless* expression in the early *Drosophila* embryo. *Development* **115**, 239-251.

Prince VE, Pickett FB. (2002) Splitting pairs: the diverging fates of duplicated genes. *Nat Rev Genet.* **3**:827-37.

Riddihough G, Ish-Horowicz D. (1991) Individual stripe regulatory elements in the *Drosophila hairy* promoter respond to maternal, gap, and pair-rule genes. *Genes Dev.* **5**:840-54.

Rivera-Pomar, R., Lu, X. G., Perrimon, N., Taubert, H. and Jackle, H. (1995). Activation of posterior gap gene expression in the *Drosophila* blastoderm. *Nature* **376**, 253-256.

Rivera-Pomar, R., and Jäckle, H. (1996). From gradients to stripes in *Drosophila* embryogenesis: filling in the gaps. *Trends Genet* **12**, 478-483.

Rivera-Pomar, R., Niessing, D., Schmidt-Ott, U., Gehring, W. J., and Jackle, H. (1996). RNA binding and translational suppression by bicoid. *Nature* **379**, 746-749.

Robertson, H. M., Preston, C. R., Phillis, R. W., Johnsonschlitz, D. M., Benz, W. K., and Engels, W. R. (1988). A stable genomic source of P-element transposase in *Drosophila melanogaster*. *Genetics* **118**, 461-470.

Robin C, Lyman RF, Long AD, Langley CH, Mackay TF. **(2002)** *hairy*: A quantitative trait locus for *drosophila* sensory bristle number. *Genetics*. **162**:155-64.

Rogers, S., Wells, R., and Rechsteiner, M. (1986). Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. *Science* **234**, 364-368.

Ronchi, E., Treisman, J., Dostatni, N., Struhl, G., and Desplan, C. (1993). Down-regulation of the *Drosophila* morphogen bicoid by the torso receptor-mediated signal transduction cascade. *Cell* **74**, 347-355.

Ronshaugen M, McGinnis N, McGinnis W. (2002) Hox protein mutation and macroevolution of the insect body plan. *Nature*. **415**:914-7.

Rubin, G. M., and Spradling, A. C. (1982). Genetic transformation of *Drosophila* with transposable element vectors. *Science* **218**, 348-353.

Rudolph KM, Liaw GJ, Daniel A, Green P, Courey AJ, Hartenstein V, Lengyel JA. (1997) Complex regulatory region mediating *tailless* expression in early embryonic patterning and brain development. *Development*. **124**:4297-308.

Rushlow C, Frasch M, Doyle H, Levine M. (1987) Maternal regulation of *zerknullt*: a homoeobox gene controlling differentiation of dorsal tissues in *Drosophila*. *Nature*. **330**:583-6

Rutherford SL, Lindquist S. (1998) Hsp90 as a capacitor for morphological evolution. *Nature*. **396**:336-42

Ruvinsky I, Ruvkun G. (2003) Functional tests of enhancer conservation between distantly related species. *Development*. **130**:5133-42. Epub 2003 Aug 27

Sachs AB, Varani G. (2000) Eukaryotic translation initiation: there are (at least) two sides to every story. *Nat Struct Biol*. **7**:356-61

Sambrook, J., Fritsch, E. F. and Maniatis, T. (1989). *Molecular cloning: a laboratory manual*. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory.

Sander K. (1975) Pattern specification in the insect embryo. *Ciba Found Symp*. **0**:241-63.

Sauer, F., Hansen, S. K. and Tjian, R. (1995a). DNA template and activator-coactivator requirements for transcriptional synergism by *Drosophila bicoid*. *Science* **270**, 1825-1828.

Sauer, F., Hansen, S. K. and Tjian, R. (1995b). Multiple TAFIIs directing synergistic activation of transcription. *Science* **270**, 1783-1788.

Sauer, F., Rivera-Pomar, R., Hoch, M. and Jäckle, H. (1996). Gene regulation in the *Drosophila* embryo. *Philos Trans R Soc Lond B Biol Sci* **351**, 579-587.

Schaeffer, V., Janody, F., Loss, C., Desplan, C. and Wimmer, E. A. (1999). Bicoid functions without its TATA-binding protein-associated factor interaction domains. *PNAS USA* **96**, 4461-4466.

Schaeffer, V., Killian, D., Desplan, C., and Wimmer, E. A. (2000). High bicoid levels render the terminal system dispensable for *Drosophila* head development. *Development* **127**, 3993-3999.

Schlötterer, C. and Tautz, D. (1992). Slippage synthesis of simple sequence DNA. *Nucleic Acids Res* **20**, 211-215.

Schmid, K. J. and Tautz, D. (1999). A comparison of homologous developmental genes from *Drosophila* and *Tribolium* reveals major differences in length and trinucleotide repeat content. *J Mol Evol* **49**, 558-566.

Schröder, R. and Sander, K. (1993). A comparison of transplantable Bicoid activity and partial Bicoid homeobox sequences in several *Drosophila* and blowfly species (Calliphoridae). *Roux's Archives of Developmental Biology* **203**, 34-43.

Schröder R, Eckert C, Wolff C, Tautz D. (2000) Conserved and divergent aspects of terminal patterning in the beetle *Tribolium castaneum*. *Proc Natl Acad Sci U S A*. **97**:6591-6

Schröder R. (2003) The genes *orthodenticle* and *hunchback* substitute for *bicoid* in the beetle *Tribolium*. *Nature*. **422**:621-5.

Schug, M. D., Mackay, T. F. and Aquadro, C. F. (1997). Low mutation rates of microsatellite loci in *Drosophila melanogaster*. *Nat Genet* **15**, 99-102.

Schug, M. D., Hutter, C. M., Wetterstrand, K. A., Gaudette, M. S., MacKay, T. F. C. and Aquadro, C. F. (1998). The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*. *Mol Biol Evol* **15**, 1751-1760.

Schulz, C., and Tautz, D. (1995). Zygotic *caudal* regulation by hunchback and its role in abdominal segment formation of the *Drosophila* embryo. *Development* **121**, 1023-1028.

Schulz, C., Schroder, R., Hausdorf, B., Wolff, C., and Tautz, D. (1998). A *caudal* homologue in the short germ band beetle *Tribolium* shows similarities to both, the *Drosophila* and vertebrate *caudal* expression patterns. *Dev Genes Evol* **208**, 283-289.

Seeger MA, Kaufman TC. (1990) Molecular analysis of the *bicoid* gene from *Drosophila pseudoobscura*: identification of conserved domains within coding and noncoding regions of the *bicoid* mRNA. *EMBO J.* **9**:2977-87.

Shashikant, C. S., Kim, C. B., Borbély, M. A., Wang, W. C. H. and Ruddle, F. H. (1998). Comparative studies on mammalian *Hoxc8* early enhancer sequence reveal a baleen whale specific deletion of a *cis*-acting element. *PNAS USA* **95**, 15446-15451.

Shaw, P. J. (1998). PhD Thesis. *Molecular characterisation of the interaction between the bicoid and hunchback genes in Musca domestica: insights into the evolution of a regulatory interaction.* In Department of Genetics. pp. 143. Leicester: University of Leicester.

Shaw, P. J., Salameh, A., McGregor, A. P., Bala, S. and Dover, G. A. (2001). Divergent structure and function of the *bicoid* gene in Muscoidea fly species. *Evol Dev* **3**, 251-262.

Shaw, P. J., Wratten, N. S., McGregor, A. P. and Dover, G.A. (2000) Co-evolution in *bicoid*-dependent promoters and the inception of regulatory incompatibilities among species of higher Diptera. *Evol Dev* **4**:265-277.

Siebert, P. D., Chenchik, A., Kellogg, D. E., Lukyanov, K. A. and Lukyanov, S. A. (1995). An improved PCR method for walking in uncloned genomic DNA. *Nucleic Acids Res* **23**, 1087-1088.

Simpson P (2002) Evolution of development in closely related species of flies and worms. *Nat Rev Genet.* **3**(12):907-17.

Simpson-Brose, M., Treisman, J. and Desplan, C. (1994). Synergy between the *hunchback* and *bicoid* morphogens is required for anterior patterning in *Drosophila*. *Cell* **78**, 855-865.

Singer JB, Harbecke R, Kusch T, Reuter R, Lengyel JA. (1996) *Drosophila brachyenteron* regulates gene activity and morphogenesis in the gut. *Development.* **122**:3707-18.

Skaer N, Simpson P. (2000) Genetic analysis of bristle loss in hybrids between *Drosophila melanogaster* and *D. simulans* provides evidence for divergence of cis-regulatory sequences in the *achaete-scute* gene complex. *Dev Biol.* **221**:148-67.

Small S, Kraut R, Hoey T, Warrior R, Levine M. (1991) Transcriptional regulation of a pair-rule stripe in *Drosophila*. *Genes Dev.* **5**:827-39.

Small, S., Blair, A. and Levine, M. (1992). Regulation of *even-skipped* stripe 2 in the *Drosophila* embryo. *EMBO J* **11**, 4047-4057.

Small S, Blair A, Levine M. (1996) Regulation of two pair-rule stripes by a single enhancer in the *Drosophila* embryo. *Dev Biol.***175**:314-24.

Sommer, R. and Tautz, D. (1991). Segmentation gene expression in the housefly *Musca domestica. Development* **113**, 419-430.

Stanojevic D, Small S, Levine M. (1991) Regulation of a segmentation stripe by overlapping activators and repressors in the *Drosophila* embryo. *Science.* **254**:1385-7.

Stauber, M., Jackle, H. and SchmidtOtt, U. (1999). The anterior determinant *bicoid* of *Drosophila* is a derived Hox class 3 gene. *PNAS USA* **96**, 3786-3789.

Stauber, M., Taubert, H. and Schmidt-Ott, U. (2000). Function of *bicoid* and *hunchback* homologs in the basal cyclorrhaphan fly *Megaselia* (Phoridae). *PNAS USA* **97**, 10844-10849.

Stauber, M., Prell, A. and Schmidt-Ott, U. (2002). A single *Hox3* gene with composite *bicoid* and *zerknüllt* expression characteristics in non-Cyclorrhaphan flies. *PNAS USA* **99**:274-9.

Stern, D. L. (1998). A role of *Ultrabithorax* in morphological differences between *Drosophila* species. *Nature* **396**, 463-466.

Stern DL. (2000) Evolutionary developmental biology and the problem of variation. *Evolution Int J Org Evolution* **54**:1079-91.

Stone JR, Wray GA. (2001) Rapid evolution of *cis*-regulatory sequences via local point mutations. *Mol Biol Evol.* **18**:1764-70.

Strecker TR, Kongsuwan K, Lengyel JA, Merriam JR. (1986) The zygotic mutant *tailless* affects the anterior and posterior ectodermal regions of the *Drosophila* embryo. *Dev Biol.* **113**:64-76

Struhl, G., Struhl, K. and Macdonald, P. M. (1989). The gradient morphogen *bicoid* is a concentration-dependent transcriptional activator. *Cell* **57**, 1259-1273.

St. Johnston, D., and Nüsslein-Volhard, C. (1992). The origin of pattern and polarity in the *Drosophila* embryo. *Cell* **68**, 201-219.

Sucena E, Stern DL. (2000) Divergence of larval morphology between *Drosophila sechellia* and its sibling species caused by *cis*-regulatory evolution of *ovo/shaven-baby. Proc Natl Acad Sci U S A.* **97**:4530-4.

Sucena E, Delon I, Jones I, Payre F, Stern DL. (2003) Regulatory evolution of *shavenbaby/ovo* underlies multiple cases of morphological parallelism. *Nature.* **424**:935-8.

Takahashi H, Mitani Y, Satoh G, Satoh N. (1999) Evolutionary alterations of the minimal promoter for notochord-specific Brachyury expression in ascidian embryos. *Development.* **126**:3725-34.

Tautz, D., Trick, M. and Dover, G. A. (1986). Cryptic simplicity in DNA is a major source of genetic variation. *Nature* **322**, 652-656.

Tautz, D., Lehmann, R., Schnürch, H., Schuh, R., Seifert, E., Kienlin, A., Jones, K. and Jäckle, H. (1987). Finger protein of novel structure encoded by *hunchback*, a 2nd member of the gap class of *Drosophila* segmentation genes. *Nature* **327**, 383-389.

Tautz, D. (1988). Regulation of the *Drosophila* segmentation gene *hunchback* by two maternal morphogenetic centres. *Nature* **332**, 281-284.

Tautz, D. and Pfeifle, C. (1989). A non-radioactive in situ hybridization method for the localization of specific RNAs in *Drosophila* embryos reveals translational control of the segmentation gene *hunchback. Chromosoma* **98**, 81-85.

Tautz, D. and Nigro, L. (1998). Microevolutionary divergence pattern of the segmentation gene *hunchback* in *Drosophila*. *Mol Biol Evol.***15**, 1403-1411.

Tautz D. (2000) Evolution of transcriptional regulation. *Curr Opin Genet Dev.* **10**:575-9.

Thisse B, Stoetzel C, Gorostiza-Thisse C, Perrin-Schmitt F. (1988) Sequence of the *twist* gene and nuclear localization of its protein in endomesodermal cells of early *Drosophila* embryos. *EMBO J.* **7**:2175-83.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-4680.

Thummel, C. S., Boulet, A. M., and Lipshitz, H. D. (1988). Vectors for *Drosophila* P-element-mediated transformation and tissue culture transfection. *Gene* **74**, 445-456.

Torigoi, E., Bennani-Baiti, I. M., Rosen, C., Gonzalez, K., Morcillo, P., Ptashne, M. and Dorsett, D. (2000). Chip interacts with diverse homeodomain proteins and potentiates *bicoid* activity *in vivo*. *PNAS USA* **97**, 2686-2691.

Treisman, J., Gonczy, P., Vashishtha, M., Harris, E., and Desplan, C. (1989). A single amino acid can determine the DNA binding specificity of homeodomain proteins. *Cell* **59**, 553-562.

Tucker-Kellogg, L., Rould, M. A., Chambers, K. A., Ades, S. E., Sauer, R. T. and Pabo, C. O. (1997). Engrailed (Gln50-->Lys) homeodomain-DNA complex at 1.9 A resolution: structural basis for enhanced affinity and altered specificity. *Structure* **5**, 1047-1054.

Umesono K, Evans RM. (1989) Determinants of target gene specificity for steroid/thyroid hormone receptors. *Cell.* **57**:1139-46

Valentine SA, Chen G, Shandala T, Fernandez J, Mische S, Saint R, Courey AJ. (1998) Dorsal-mediated repression requires the formation of a multiprotein repression complex at the ventral silencer. *Mol Cell Biol.* **18**:6584-94.

Waddington, C. H. (1942). Canalization of development and the inheritance of acquired characters. *Nature* **150**, 563-565.

Wang RL, Stec A, Hey J, Lukens L, Doebley J. (1999) The limits of selection during maize domestication. *Nature*. **398**:236-9.

White LD, Coates CJ, Atkinson PW, O'Brochta DA. (1996) An eye color gene for the detection of transgenic non-drosophilid insects. *Insect Biochem Mol Biol.* **26**:641-4

Wilkins AS. (1997) Canalization: a molecular genetic perspective. *Bioessays.* **19**:257-62.

Wimmer, E. A., Simpson-Brose, M., Cohen, S. M., Desplan, C., and Jäckle, H. (1995). *Trans*-acting and *cis*-acting requirements for blastoderm expression of the head gap gene *buttonhead*. *Mech Dev* **53**, 235-245.

Wimmer, E. A., Carleton, A., Harjes, P., Turner, T. and Desplan, C. (2000). Bicoid-independent formation of thoracic segments in *Drosophila*. *Science* **287**, 2476-2479.

Wittkopp PJ, Vaccaro K, Carroll SB. (2002) Evolution of *yellow* gene regulation and pigmentation in *Drosophila*. *Curr Biol.* **12**:1547-56.

Wolff, C., Schroder, R., Schulz, C., Tautz, D. and Klingler, M. (1998). Regulation of the *Tribolium* homologues of *caudal* and *hunchback* in *Drosophila*: evidence for maternal gradient systems in a short germ embryo. *Development* **125**, 3645-3654.

Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA. (2003) The evolution of transcriptional regulation in eukaryotes. *Mol Biol Evol.* **20**:1377-419.

Wu LH, Lengyel JA. (1998) Role of *caudal* in hindgut specification and gastrulation suggests homology between *Drosophila* amnioproctodeal invagination and vertebrate blastopore. *Development.* **125**:2433-42.

Xu X, Xu PX, Suzuki Y. (1994) A maternal homeobox gene, *Bombyx caudal*, forms both mRNA and protein concentration gradients spanning anteroposterior axis during gastrulation. *Development.* **120**:277-85.

Xu PX, Zhang X, Heaney S, Yoon A, Michelson AM, Maas RL. (1999) Regulation of *Pax6* expression is conserved between mice and flies. *Development.* **126**:383-95.

Yu RT, McKeown M, Evans RM, Umesono K. (1994) Relationship between *Drosophila* gap gene *tailless* and a vertebrate nuclear receptor Tlx. *Nature.* **370**:375-9.

Yuan, D., Ma, X. and Ma, J. (1996). Sequences outside the homeodomain of bicoid are required for protein- protein interaction. *J Biol Chem* **271**, 21660-21665.

Yuan, D., Ma, X. and Ma, J. (1999). Recognition of multiple patterns of DNA sites by *Drosophila* homeodomain protein Bicoid. *J Biochem* (Tokyo) **125**, 809-817.

Zakany J, Duboule D. (1999) *Hox* genes in digit development and evolution. *Cell Tissue Res.* **296**:19-25.

Zakany J, Fromental-Ramain C, Warot X, Duboule D. (1997) Regulation of number and size of digits by posterior *Hox* genes: a dose-dependent mechanism with potential evolutionary implications. *Proc Natl Acad Sci U S A.* **94**:13695-700

Zhao, C., Dave, V., Yang, F., Scarborough, T. and Ma, J. (2000). Target selectivity of bicoid is dependent on nonconsensus site recognition and protein-protein interaction. *Mol Cell Biol* **20**, 8112-8123.

Zhao C, York A, Yang F, Forsthoefel DJ, Dave V, Fu D, Zhang D, Corado MS, Small S, Seeger MA, Ma J. (2002) The activity of the *Drosophila* morphogenetic protein Bicoid is inhibited by a domain located outside its homeodomain *Development* **129**:1669-80.

Zhu, W. and Hanes, S. D. (2000). Identification of *Drosophila* Bicoid-interacting proteins using a custom two-hybrid selection. *Gene* **245**, 329-339.

Zhu, W. C., Foehr, M., Jaynes, J. B. and Hanes, S. D. (2001). *Drosophila* SAP18, a member of the Sin3/Rpd3 histone deacetylase complex, interacts with Bicoid and inhibits its activity. *Dev Genes Evol* **211**, 109-117.