High Sequence Turnover in the Regulatory Regions of the Developmental Gene hunchback in Insects

J. M. Hancock,* P. J. Shaw,† F. Bonneton,‡ and G. A. Dover†

*MRC Clinical Sciences Centre, Gene and Genome Evolution Group, Hammersmith Hospital, London, England; and †Department of Genetics, University of Leicester, Leicester, England

Extensive sequence analysis of the developmental gene hunchback and its 5’ and 3’ regulatory regions in Drosophila melanogaster, Drosophila virilis, Musca domestica, and Tribolium castaneum, using a variety of computer algorithms, reveals regions of high sequence simplicity probably generated by slippage-like mechanisms of turnover. No regions are entirely refractory to the action of slippage, although the density and composition of simple sequence motifs varies from region to region. Interestingly, the 5’ and 3’ flanking regions share short repetitive motifs despite their separation by the gene itself, and the motifs are different in composition from those in the exons and introns. Furthermore, there are high levels of conservation of motifs in equivalent orthologous regions. Detailed sequence analysis of the P2 promoter and DNA footprinting assays reveal that the number, orientation, sequence, spacing, and protein-binding affinities of the BICOID-binding sites varies between species and that the ‘P2’ promoter, the nanos response element in the 3’ untranslated region, and several conserved boxes of sequence in the gene (e.g., the two zinc-finger regions) are surrounded by cryptically-simple-sequence DNA. We argue that high sequence turnover and genetic redundancy permit both the general maintenance of promoter functions through the establishment of coevolutionary (compensatory) changes in cis- and trans-acting genetic elements and, at the same time, the possibility of subtle changes in the regulation of hunchback in the different species.

Introduction

Eukaryotic genomes are subject to a number of non-Mendelian mechanisms of sequence turnover (transposition, unequal crossing over, retrotransposition, slippage, gene conversion, etc.), generating high levels of redundancy and sequence divergence within and between species (for reviews see Dover 1982, 1993; Tautz, Trick, and Dover 1986; Hancock 1996a, 1996b). How are essential biological functions maintained during such turmoil? One solution is for natural selection to ensure the sequence and structural integrity of critically important genomic regions through the selective elimination of unwanted variants. Hence, searching for long conserved sequences across species is considered a first step in the location and dissection of functionally important regions in both genes and gene regulatory elements. One consequence of this approach is that there is a working assumption that highly diverged sequences are neutral in function and free to drift and decay. However, diverged sequences may harbor important functions that have been maintained despite divergence. Such functional maintenance is made possible by the process of “molecular coevolution,” particularly since it operates among multiple-copy genetic elements (Dover and Flavell 1984; Dover 1992).

Essentially, molecular coevolution depends on natural selection promoting the spread of compensatory mutations in genes whose products functionally interact with redundant genetic elements subject to high levels of turnover and divergence. The phenomenon was first recognized in multiple promoters of the rDNA multigene families. For several genera of plants and animals, it has been observed that the poll transcription machinery of one species is incapable of efficiently recognizing the promoter sequences of another, often closely related, species (for references, see Dover and Flavell 1984; Dover 1992; Evers and Grummt 1995). One would expect, given the critical role of ribosomal RNA in the translation machinery, that there would be a high premium for retention of a good working relationship between rDNA promoters and terminators and their respective complexes of proteins that are required to activate them. Instead, the continual differentiation of multiple promoters and terminators, known in these cases to be subject to turnover by unequal crossing over and gene conversion, has led to the coevolution of protein cofactors, such that essential functions are maintained within a species while intertaxon differences accumulate. Molecular coevolution is an outcome of the interaction of the homogenization processes of turnover (molecular drive) and natural selection; it has subsequently been observed between segments of ribosomal RNAs that accumulate novel sequences by slippage (Hancock and Dover 1990; Hancock 1995a) and may also contribute in a number of ways to protein evolution (Hancock 1996b).

Is molecular coevolution a feature of promoters and their transcriptional complexes in genes other than the highly redundant rDNA genes? To investigate this question, we chose to dissect the genetic and functional differences between species of higher diptera, in particular between Drosophila and Musca with regard to the interaction between bicoid and hunchback, two genes involved in the determination of anterior-posterior polarity in early development. The maternally expressed bicoid gene produces a homeodomain protein that recognizes a number of BICOID-binding sites in the proximal P2 promoter of the gap gene hunchback, an interaction that...
is essential for the correct expression of *hunchback* in the anterior half of the egg (reviewed in St. Johnston and Nüsslein-Volhard 1992). Our recent results on the *bicoid–hunchback* interaction in the housefly *Musca domestica* (Bonneton et al. 1997) indicate that there are important sequence, spacing, orientation, and copy number differences between the two species regarding the BICOID-binding sites, which might explain the in vivo differences we have observed in *hunchback* expression patterns and the in vitro differences in the strengths of binding of homeodomains to the BICOID-binding sites in heterologous mixtures between the two species (Bonneton et al. 1997; unpublished data). Additionally, there are high levels of amino acid differences (6 out of 60) in the BICOID homeodomain between the two species (Sommer and Tautz 1991), which might reflect the co-evolution of compensatory changes between homeodomains and their binding sites.

In order to fully understand the mutational forces at work on the *hunchback* promoter, it is important to reveal the processes of turnover both among and surrounding the BICOID-binding sites. How and why did the binding sites alter in number and sequence between the species? Is the P2 promoter an island of relative stability in a sea of sequence turbulence flanking the 5' end of the *hunchback* gene? Similarly, are the functionally important sites in the 3' flanks of *hunchback* (e.g., the nanos response element, [NRE]) involved with RNA stability and localization (Wharton and Struhl 1991) relatively stable during sequence turnover? To answer these important questions, we present in this paper the results of sequence analysis and some functional tests of the *hunchback* gene and its 5' and 3' flanking regions from a number of species. Our approaches show that there are high levels of slippage-dominated turnover operating across the *hunchback* regions, involving the coding regions, the 5' promoters, and 3' regulatory signals, and that the slippage-generated short repetitive motifs can be shared extensively both horizontally across the regions and vertically between species. Hence, notwithstanding the unalignability of, for example, the promoter regions between the species in question, there are interesting levels of shared structural features, involving the composition and location of slippage-generated simple-sequence DNA. Such features could be of functional significance, a possibility that is open for testing. Specifically, although functionally important regions such as the BICOID-binding sites, the NREs, and the zinc-finger domains of *hunchback* are not quite as in turmoil as the surrounding sequences, they are nevertheless subject to turnover, resulting in variation in copy number, orientation, location, sequence, and protein-binding affinity.

### Materials and Methods

The sequences used for this analysis were *M. domestica* (GenBank entry Y13050; Bonneton et al. 1997), *Drosophila melanogaster* (GenBank entry Y00274; Tautz et al. 1987), *Drosophila virilis* (GenBank entry X15359; Treier, Pfeifle, and Tautz 1989), and *Tribolium castaneum* (GenBank entry X91618; Wolff et al. 1995).

Dot matrix analysis of sequences was carried out using DottyPlot (Gilbert 1990) using a stringency of 19-base perfect match in a window of 35 bp (Hancock and Dover 1988). Displays of distribution of individual triplet motifs were generated using DottyPlot by comparing sequences against a file containing an array of all 64 possible triplet motifs (Hancock and Dover 1990).

Sequence simplicity analysis was carried out using SIMPLE34 (Hancock and Armstrong 1994), a modified version of Tautz, Trick, and Dover’s (1986) SIMPLE program. This program investigates the distribution of tri- and tetranucleotide repeats within a sequence by counting the occurrences of each motif in the sequence within a window 32 bp 5' and 3' to it. The window is scored for occurrences of the tri- and tetranucleotide motifs at its center by awarding four points for each occurrence of the tetranucleotide motif and three points for each trinucleotide (Tautz, Trick, and Dover 1986). Sequence simplicity plots reflect the distribution of these scores along the sequence, with peaks reflecting regions of high concentrations of particular motifs. The program calculates the frequency distribution of the scores generated by all 64-bp windows and compares this distribution with that for 10 random sequences of the same length and base doublet composition. A probability that any given score occurs by chance can then be calculated by comparing these frequency distributions. Motifs associated with windows whose scores occur at frequencies that exceed the average frequency for that score in the random sequences by a factor of 10 or more are then designated significantly simple motifs (SSMs), motifs that are clustered at particular positions within the sequence (Hancock and Armstrong 1994). This designation does not represent a measure of statistical significance of motif clustering, but serves to identify sequence motifs associated with peaks of sequence simplicity as measured by the program.

Motifs tandemly repeated more than five times within the sequence were identified using ARRAYFINDER, a FORTRAN77 program written on a Sun SparcStation 5. The program scans along a sequence and at each point identifies the longest tandem repeat with a basic motif less than or equal to 10 bases that starts at that point. The basic motif and motif copy number of this longest repeat are recorded and subsequently analyzed by a second program, READARRAYFILES, which analyzes the frequency and length distribution of each repeated motif.

Multiple alignment of protein-coding sequences was carried out using PILEUP. Alignment editing was carried out using LINEUP, and similarity profiles were calculated using PLOTSIMILARITY. All of these programs are part of the GCG (1994) package (version 8.1).

 Parsimony analysis of BICOID binding sites was carried out using the DNAPARS module of version 3.572 of the PHYLIP package for Power Macintosh computers (Felsenstein 1993). The consensus of the most-parsimonious trees was generated using the CONSENSE module.

DNase I footprinting autoradiographs from our previous work (Bonneton et al. 1997) were scanned, and
the images were analyzed on a Power Macintosh model 4400/200 computer using the public domain NIH image program (developed at the US NIH and available on the internet at http://rab.info.nih.gov/nih-image). DNase I footprinting involved the Musca BICOID homeodomain fusion protein and the 1 kb of hunchback promoter. Data for sense and antisense strand protection are averaged from two experiments for each, using the same binding conditions in which BICOID homeodomain was in excess of DNA. Fractional occupancy was determined by comparing the band intensities of the homeodomain-bound lanes with those of the control lanes using the following equation, as described by Brenowitz et al. (1986).

\[
Y_{\text{app}} = 1 - \frac{(DN_{\text{site}}/DN_{\text{std}})}{(DR_{\text{site}}/DR_{\text{std}})},
\]

where \(Y_{\text{app}}\) is fractional occupancy; \(DN_{\text{site}}\) is average band intensity of bound region in BICOID lanes (\(N\)); \(DN_{\text{std}}\) is average band intensity of standard block outside bound region in lanes (\(N\)); \(DR_{\text{site}}\) is average band intensity of bound region in control lanes, no BICOID added (\(R\)); and \(DR_{\text{std}}\) is average band intensity of standard block outside bound region in lanes (\(R\)).

A block of bands corresponding to the 5-bp core sequence 5′-TAATC-3′ common to the BICOID-binding consensus sequences for the species studied (Bonneton et al. 1997) was chosen as the bound region for each binding site. For further details, see the footnote to table 1.

### Results

Our results are displayed in a number of figures in which four species are analyzed, the three dipterans *M. domestica*, *D. melanogaster*, and *D. virilis* and the beetle *T. castaneum*. The sources of the sequences, including our own of *M. domestica*, are given in Materials and Methods.

Within-Species Fine-Grained Sequence Redundancy Revealed by Dot Matrix Analysis

Figure 1 (top panels) shows the results of a dot matrix analysis involving up to 8 kb of sequence from *M. domestica* and *T. castaneum*. To save space, we do not show the very similar dot matrix patterns for the two *Drosophila* species. At this level of resolution, it is apparent that in all four species, albeit differing extents, there are two main regions (the flanking regions and the gene) that are composed of short repetitive motifs and are shared mainly within but not between regions. For example, the 5′ and 3′ flanking regions of *M. domestica* contain densely spaced motifs that are shared. The two exons and the intron separating them also contain short repetitive motifs that are largely of different composition and frequency than the 5′ and 3′ flanking regions. A similar pattern of sharing of densely spaced short repetitive motifs can be seen in *T. castaneum*, with the regions composed of different motifs.

What Precisely Are the Short Repetitive Motifs in the Sequences?

To identify the short repetitive motifs in the sequences, we applied an alternative analysis (Hancock and Dover 1990), as depicted in figure 1 (middle panels). The displays have 64 horizontal lines, each representing one trinucleotide motif, from AAA at the top to TTT at the bottom. Each small vertical line represents a given position of a particular triplet in the overall sequence. The panels at the bottom show the frequencies of different tandem repeats that are more than five motifs long in each of the functional units.

From the triplet displays, it is possible to trace by eye the occurrence of any given triplet across the sequence. For example, the 5′ and 3′ flanking regions of *M. domestica* contain about a dozen prominent triplets running throughout the sequences. In the region of the P2 promoter, there is a more widespread occurrence of triplets of different composition. Similarly, the gene region itself contains runs of triplet motifs, some of which are specific to this region and are relatively more densely packed than in other regions.

The display of triplets in *T. castaneum* also reveals the occurrence of runs of triplets which are shared by 5′ and 3′ flanking regions to the exclusion of the gene region, which has a more even spread of the available triplets. In contrast, the gene region of *D. virilis* (not shown) is more densely packed with other specific triplets, supporting the results of the dot matrix. In general, these triplets are different from those shared by the 5′ and 3′ flanking regions of *D. virilis*. *Drosophila melanogaster* has a more even occurrence and spread of triplets, excepting some prominent runs in the gene regions.

Poly (A) and poly (T) are by far the most common tandem repeats throughout all the species, as has been found for a number of other genomes (Hancock 1995b, 1996c). CAA and CAT repeats appear in the *M. domestica* gene region, as does AGC (=CAG; known as opa repeats; Wharton et al. 1985) in *D. melanogaster*. The *D. virilis* gene region contains long tandem arrays of CAA, CAT, and CAG. *Tribolium castaneum* has only mononucleotide repeats throughout the regions, with C and G repeats prominent in the gene region.

### Table 1

**Musca domestica** BICOID-Binding Site Strength Data

<table>
<thead>
<tr>
<th>Binding Site</th>
<th>Antisense Fractional Occupancy</th>
<th>Sense Fractional Occupancy</th>
<th>Strand-Averaged Fractional Occupancy</th>
<th>Binding Strength</th>
</tr>
</thead>
<tbody>
<tr>
<td>md A</td>
<td>0.84</td>
<td>0.55</td>
<td>0.70</td>
<td>W</td>
</tr>
<tr>
<td>md B</td>
<td>0.53</td>
<td>0.73</td>
<td>0.63</td>
<td>W</td>
</tr>
<tr>
<td>md C</td>
<td>0.86</td>
<td>0.94</td>
<td>0.90</td>
<td>S</td>
</tr>
<tr>
<td>md D</td>
<td>0.46</td>
<td>0.86</td>
<td>0.66</td>
<td>W</td>
</tr>
<tr>
<td>md E</td>
<td>0.95</td>
<td>0.90</td>
<td>0.93</td>
<td>S</td>
</tr>
<tr>
<td>md F</td>
<td>0.92</td>
<td>0.88</td>
<td>0.90</td>
<td>S</td>
</tr>
<tr>
<td>md G</td>
<td>0.94</td>
<td>0.81</td>
<td>0.87</td>
<td>S</td>
</tr>
<tr>
<td>md H</td>
<td>0.75</td>
<td>0.73</td>
<td>0.74</td>
<td>W</td>
</tr>
<tr>
<td>md I</td>
<td>0.92</td>
<td>0.97</td>
<td>0.95</td>
<td>S</td>
</tr>
<tr>
<td>md J</td>
<td>0.81</td>
<td>0.99</td>
<td>0.90</td>
<td>S</td>
</tr>
</tbody>
</table>

**NOTE.**—Binding sites are classified arbitrarily as weak or strong. Weak sites have strand-averaged fractional occupancies of \(\leq 0.75\), whereas strong sites have strand-averaged fractional occupancies of \(\geq 0.85\).

Within-Species Fine-Grained Sequence Redundancy Revealed by Dot Matrix Analysis

Figure 1 (top panels) shows the results of a dot matrix analysis involving up to 8 kb of sequence from *M. domestica* and *T. castaneum*. To save space, we do not show the very similar dot matrix patterns for the two *Drosophila* species. At this level of resolution, it is apparent that in all four species, albeit differing extents, there are two main regions (the flanking regions and the gene) that are composed of short repetitive motifs and are shared mainly within but not between regions. For example, the 5′ and 3′ flanking regions of *M. domestica* contain densely spaced motifs that are shared. The two exons and the intron separating them also contain short repetitive motifs that are largely of different composition and frequency than the 5′ and 3′ flanking regions. A similar pattern of sharing of densely spaced short repetitive motifs can be seen in *T. castaneum*, with the regions composed of different motifs.

What Precisely Are the Short Repetitive Motifs in the Sequences?

To identify the short repetitive motifs in the sequences, we applied an alternative analysis (Hancock and Dover 1990), as depicted in figure 1 (middle panels). The displays have 64 horizontal lines, each representing one trinucleotide motif, from AAA at the top to TTT at the bottom. Each small vertical line represents a given position of a particular triplet in the overall sequence. The panels at the bottom show the frequencies of different tandem repeats that are more than five motifs long in each of the functional units.

From the triplet displays, it is possible to trace by eye the occurrence of any given triplet across the sequence. For example, the 5′ and 3′ flanking regions of *M. domestica* contain about a dozen prominent triplets running throughout the sequences. In the region of the P2 promoter, there is a more widespread occurrence of triplets of different composition. Similarly, the gene region itself contains runs of triplet motifs, some of which are specific to this region and are relatively more densely packed than in other regions.

The display of triplets in *T. castaneum* also reveals the occurrence of runs of triplets which are shared by 5′ and 3′ flanking regions to the exclusion of the gene region, which has a more even spread of the available triplets. In contrast, the gene region of *D. virilis* (not shown) is more densely packed with other specific triplets, supporting the results of the dot matrix. In general, these triplets are different from those shared by the 5′ and 3′ flanking regions of *D. virilis*. *Drosophila melanogaster* has a more even occurrence and spread of triplets, excepting some prominent runs in the gene regions.

Poly (A) and poly (T) are by far the most common tandem repeats throughout all the species, as has been found for a number of other genomes (Hancock 1995b, 1996c). CAA and CAT repeats appear in the *M. domestica* gene region, as does AGC (=CAG; known as opa repeats; Wharton et al. 1985) in *D. melanogaster*. The *D. virilis* gene region contains long tandem arrays of CAA, CAT, and CAG. *Tribolium castaneum* has only mononucleotide repeats throughout the regions, with C and G repeats prominent in the gene region.
Figure 1.—Patterns of sequence repetition within hunchback gene regions of *M. domestica* and *T. castaneum*. Top, Dot matrix comparison of the complete gene region against itself, using a match stringency of 19 out of 35 (Hancock and Dover 1988). Scales along the top and right-hand side represent sequence coordinates. Below is a cartoon map of the region. Gray boxes represent non-protein-coding exonic sequences, black boxes represent protein-coding exonic sequences, and hatched boxes represent zinc-finger-coding regions. BICOID-binding sites are indicated by vertical lines, and NREs, where known, are indicated by downward-pointing arrows. Middle, The distribution of all possible triplet motifs within the sequence. The scale along the top represents position within the sequence. Boxes along the left-hand side allow identification of individual motifs. Motifs are arranged alphabetically, from AAA (top) to TTT (bottom). Bottom, Frequencies of occurrence of tandem repeats of length five or greater for each subregion as defined for figure 2, estimated using ARRAYFINDER (see Materials and Methods). Note that
Cryptic Simplicity

Are the patches of high cross-matching redundancy revealed in the dot plots a true reflection of sequence composition of the regions being tested? To answer this question, we analyzed the sequences using the SIMPLE34 program (Hancock and Armstrong 1994; see Materials and Methods) based on the method originally described in Tautz, Trick, and Dover (1986). Briefly, the method reveals statistically significant regions of simple-sequence DNA as it occurs either in tandemly repeated arrays (which we termed pure simple-sequence DNA in Tautz, Trick, and Dover [1986]) or in regions in which a few repetitive motifs are scrambled (cryptic simplicity). Such cryptically simple DNA would not be distinguishable by the naked eye from randomly dispersed nucleotides. The graphic displays (fig. 2) of high "mountain peaks" indicate that almost all regions are cryptically simple through the repetition of one short motif or another. The motifs associated with these peaks (which we have termed SSMs; Hancock and Armstrong 1994; see Materials and Methods) are indicated. Some regions can contain a number of SSMs; such regions would represent cryptically simple regions. Other regions consist of one predominant SSM; such regions probably contain tandem arrays of a single SSM.

It is interesting to note that the sequences of T. castaneum generally have lower levels of cryptically significant regions, with no detectable SSMs. The dot plot of this species, however, does reveal dense levels of short-motif repetition. This seeming contradiction indicates that a sequence can be composed of many different short motifs that are shared across the regions, but that no one motif is particularly abundant. The dot matrix analysis and the cryptic simplicity profiles and their attendant SSMs need to be considered together for a full interpretation of the nature of the repetitive structure in a given sequence.

It is clear from figures 1 and 2 that specific motifs occur in extensive regions, that different regions can have different specific repeats, and that nonadjacent regions can share the same specific motifs. How and why are such patterns generated? If slippage or slippage-like processes are capable of generating pure or cryptically-simple-sequence DNA of any composition, how have such processes produced the same repetitive short motifs over such extensive areas and sometimes in widely separated regions? These issues and similar findings for other genomic regions are discussed later.

Are the Repetitive Motifs Shared Between Species?

Figures 1 and 2 show, by a variety of analytical methods, that within any given species, short repetitive motifs are abundant and widely shared horizontally across the analyzed sequences of the hunchback gene and its 5' and 3' flanking regions. Are such repetitive structures shared vertically between the species? To answer this question, we made interspecific dot matrices between M. domestica and each of the three other species and between the two Drosophila species. In figure 3, we show just two of these comparisons: M. domestica versus T. castaneum, and M. domestica versus D. virilis. It is apparent, particularly between M. domestica and each of the three other species, that there is extensive cross-matching in both the 5' and 3' flanking regions and in the gene region, with a lesser degree of cross-matching between the two Drosophila species (data not shown). Indeed, it would be hard to distinguish the within-species matrices from the between-species matrices.

How and why has such extensive interspecific sharing occurred? Is this a result of selective conservation of sequence structure for functional purposes, or is there an element of coincidental occurrences of short motifs, given the AT/GC composition of the sequence substrate on which slippage-like processes are operating and the mechanistic biases of slippage itself?

Fine Detail of the BICOID-Binding and NRE Regulatory Regions

In order to answer these questions, it is necessary to home in on the details of variation within and around the BICOID-binding sites and the 3' untranslated region (UTR) regulatory regions. Figure 4a and b show the cryptic simplicity profiles around the P2 promoters of M. domestica, D. melanogaster, and D. virilis and around the 3' UTRs of all four species. It can be seen that the BICOID-binding sites and the NREs are in regions of cryptically simple DNA with some prominent nearby landmarks of both tandemly arrayed and interspersed (cryptic) repeats (see legend to fig. 4 for details). Application of the SIMPLE 34 program (see Materials and Methods) to just the region containing the BICOID-binding sites yielded relative simplicity factors (RSFs) of 1.628 for Musca (coordinates 2626–3210) (P < 0.003), 1.228 for D. virilis (4951–5237) (NS [>0.05]), and 1.502 for D. melanogaster (3906–4228) (P < 0.01). Thus, even the narrow regions containing the BICOID-binding sites are significantly simple in Musca and D. melanogaster. This is not the case for D. virilis, but this might reflect the high overall repetitiveness of the D. virilis sequence, which could lead to the random control sequences also being repetitive.

In our earlier paper (Bonneton et al. 1997), we described the increases in copy number and intersite spacings of M. domestica BICOID-binding sites relative to those of D. melanogaster. Close inspection of the BICOID-binding sites of D. melanogaster and M. domestica provides hints as to how new sites might have arisen. For example, in M. domestica, the spacings between sites A–C and D–F are similar, which may reflect a duplication by a recombination event. Second, M. domestica sites I and J are oriented tail-to-tail and are separated by a region in which a (GT)n motif is followed by

---

scales differ between subregions. For T. castaneum, subregions E2 and E3 are the coding exons, interrupted by the intron subregion I2. No significant tandem repeats were found in the third untranslated exon or the short adjacent intron preceding the coding region for T. castaneum.
**Drosophila melanogaster**

**Drosophila virilis**

**Musca domestica**

**Tribolium castaneum**

---

**Fig. 2.**—Sequence simplicity within *hunchback* gene regions for the four indicated species; smoothed average plot (window size 10) of simplicity scores achieved by individual 64-bp windows during simplicity analysis using SIMPLE34 (Hancock and Armstrong 1994). Labels above peaks represent the SSMs associated with each peak (see Materials and Methods for explanation). Cartoon maps of the regions are shown as in figure 1.
Figure 3.—Comparison of sequence repetition patterns between pairs of species as indicated. Each panel is a dot plot comparing the two sequences at stringency 19/35. Scales along the top and the right-hand side represent sequence position. Maps of each gene region (see fig. 1) are presented for each sequence.

a (CA)$_4$ motif. Could these sites have arisen by formation of a foldback loop? Third, in D. melanogaster, the weak BICOID sites (broken arrows in fig. 4A) are generally associated with regions of relatively high simplicity, especially for weak site x4 (described by Ma et al. 1996), which is in the middle of a cryptically simple region based on the ATCC motif. This site appears to be derived from the preceding strong site (A2). Other weak sites in the three species are not necessarily associated with cryptic simplicity. In general, the BICOID sites are very AT-rich and are surrounded by GC-rich bases. The NREs, on the other hand, are GC-rich, with some prominent runs of T that seem specific. (For further details, see figure 4B and the discussion below.)

Comparisons of Binding-Site Affinities Within Promoter Modules

The data presented in table 1 show relative strengths of DNA binding for sites in the M. domestica promoter derived from the DNase I footprinting data published in Bonneton et al. (1997). For the weak binding sites A, B, and H, protection on the antisense strand is noticeably much weaker than that on the sense strand. In contrast, for the strong sites C, E, G, I, and J, the protections on the sense and antisense strands are similar. DNase I footprinting studies of the D. melanogaster promoter have shown that for the weak sites x1, x2, and x3, protection is observed only on the antisense strand, whereas for the sense strand, hypersensitive bands in the BICOID-bound lanes are observed (Driever and Nüsslein-Volhard 1989). For the weak site x4, protein interaction is detected only as hypersensitive bands in the BICOID-bound lanes (Ma et al. 1996).

Detailed footprint titrations have shown that the differences in affinity between weak and strong sites in isolation are less than previously thought, and in the natural promoter, weak sites are bound in a cooperative fashion similar to strong sites (Ma et al. 1996). The weak BICOID-binding sites in D. melanogaster and M. domestica and the weak site x3 inferred from sequence alignment in D. virilis (Lukowitz et al. 1994) all contain mismatches to the 5′-TAATC-3′ core common to the consensus binding sequence in all three species (with the exception of site E in M. domestica). Quantitative in vitro studies of the lysine-50 variant engrailed homeodomain (BICOID-like specificity) bound to different artificial sites have shown the effects of changing different positions in the 5 bp core (Ades and Sauer 1995). In the main, the core mismatches in the weak sites occur at positions of less functional significance compared with other positions shown to be importantly involved with sequence recognition in structural studies of DNA-homeodomain crystals (Tucker-Kellogg et al. 1997).

What Are the Evolutionary Relationships Between BICOID-Binding Sites?

In figure 5A, we align the binding-site sequences (a 14-bp region that includes the 9-bp consensus defined
by Driever and Nüsslein-Volhard (1989) of *M. domestica* and the two *Drosophila* species and depict a tree of relationships based on these sequences. The BICOID sites cluster into two large, well-defined families which we call family A1 and family A2, because they contain the strong BICOID-binding A1 and A2 sites of *D. melanogaster* (and their *D. virilis* homologs), respectively. Interestingly, the *M. domestica* and *Drosophila* sites are distributed between the two main families and also in positions outside of the families: hence, the sites are not phylogenetically clustered according to species. The weak BICOID-binding sites in *D. melanogaster* are located at the margins of the two families in the tree. The sequence proximity of sites within any given family might reflect either common origins or evolutionary convergence. The genomic clustering of particular A1,
Regulatory Evolution of a Developmental Gene

A

A1 Family

A2 Family

B

Musca domestica

Drosophila melanogaster

Drosophila virilis
A2, or other pairs of sites in the P2 promoter in *M. domestica* and *D. melanogaster* (see legend to fig. 5B), albeit sometimes with different relative orientations, suggests that local processes of duplication or deletion might have been involved in the history of the distribution of sites. Such duplications and some subsequent divergence in sequence might also have changed the binding affinities of sites. Similarly, the weak site x3 (dotted circle in fig. 5B) in *D. virilis* is physically close to site A1 (striped circle), suggestive of a duplication. However, site x3 in *D. virilis* is close in sequence to site x3 in *D. melanogaster* in family A1 (fig. 5A), indicating that the possible duplication that gave rise to A1 and x3 arose prior to species divergence. Interestingly, the strong site A3 (shaded box) in *D. melanogaster* is also present in a similar position in *D. virilis*.

However, *M. domestica* sites B, C, and G (black boxes) lie outside the two families but do not relate closely in sequence to the *Drosophila* A3 site. All sites either arose independently, or their origins are now obscured by point mutation. The close relationship between *M. domestica* sites B and C might indicate an origin by duplication after they diverged from site G.

**Are the Protein-Coding Sequences Refractory to Sequence Turnover?**

In figure 6, we show the simplicity profiles of the *hunchback* genes of the four species. From amino acid sequence alignments (not shown), relatively conserved blocks of sequence are indicated as boxes. Two previously identified conserved blocks in higher dipterans (the C and D boxes) described by Hülskamp et al. (1994) are shown, as are five boxes, A–E, shared by all four species but of unknown function. Outside of such conserved regions, it is clear that high levels of cryptic simplicity do reside in exonic sequences, and the compositions of the triplet repeats and other tandem arrays in such regions are presented in figure 1.

**Discussion**

There is much current interest in the structure and evolution of the genetic elements regulating gene expression and in the experimental dissection of the forces that have shaped their history (for reviews of literature and experimental data, see Kreitman and Ludwig 1996; Arnone and Davidson 1997; Bonneton et al. 1997; Ludwig, Patel, and Kreitman 1998). Eukaryotic promoters consist of modular and redundant elements that are bound by a number of trans-acting regulatory proteins and have been shown to vary in copy number, sequence, interelement spacing, binding affinity, and orientation within and between species in some well-studied cases. Cooperativeness and synergy between regulatory proteins and between binding sites varying in the above parameters ensure that the overall consequences of mutational changes on gene expression are not necessarily an all-or-nothing effect. Instead, there is a degree of robustness emerging from complex promoters that permits subtle evolutionary changes to take place while maintaining the overall promoter function. This dual property of maintenance and change is reinforced if there are also compensatory mutational changes occurring in the genes coding for the regulatory proteins in response to changes in promoter structure and sequence. Theoretical modeling of the BICOID *hunchback* interaction has suggested that the overall function of transcriptional promotion can be maintained through compensatory mechanisms acting in cis and trans (Gibson 1996). A general feature of developmental genes is that extensive buffering is present through overlapping functions, such that there is inherent plasticity in the system (Tautz 1992).

The ability of coevolutionary changes to occur between promoters and their transacting regulatory proteins is facilitated by the redundancy of 5′ binding sites and the high levels of sequence turnover among the sites. Such compensatory changes permit the maintenance of the general functions of the promoter while at the same time allowing subtle taxon-specific differences to accumulate (Dover and Flavell 1984; Dover 1992).

In order to understand the extent of molecular coevolution and the mechanisms underlying molecular co-
evolution in promoters, we embarked on a molecular genetic comparison of the \textit{hunchback} promoter and its multiple BICOID-binding sites in higher diptera—in particular, between \textit{Drosophila} spp. and the house fly \textit{M. domestica}. In our earlier paper (Bonneton et al. 1997), we described the interspecific differences in the structure of the P2 promoter and examined the differences in promoter function in transgenic manipulations. From these studies, we could conclude that the \textit{Musca} promoter cannot completely substitute the \textit{D. melanogaster} promoter function during early and late embryogenesis and that such differences might be a consequence of the differences in sequence, number, orientation, and spacing of BICOID-binding sites in conjunction with the large number of amino acid differences within the homeodomain BICOID between the two genera. In order to understand the genomic forces responsible for changes in promoter organization and their effects on promoter function, we performed a detailed comparative analysis of sequence turnover in \textit{hunchback} and its 5' and 3' flanking regions and compared protein-binding-site affinities and functions in in vitro and in vivo assays. The data on the fine structural analysis of sequence turnover and binding-site protein affinities are presented in this paper, and the extensive results of the functional assays are to be presented elsewhere (unpublished data).

From the analysis of the \textit{hunchback} regions in \textit{D. melanogaster}, \textit{D. virilis}, \textit{M. domestica} and \textit{T. castaneum}, we make the following observations:

- Extensive regions of the \textit{hunchback} locus, including 5' and 3' flanking regions with their regulatory elements and the \textit{hunchback} gene itself, contain cryptically simple-sequence DNA.
- The particular short repetitive motifs that make up the simple-sequence DNA can be shared extensively for several kilobases of sequence, and between 5' and 3' flanking regions, separated by the gene.
- The coding sequences are also cryptically simple for short repetitive motifs that are not generally shared with the 5' and 3' flanking regions.
- The densities of short repetitive motifs can vary, depending on the functional constraints on the sequence in question. For example, the P2 promoters, the 3' NREs, and several conserved elements within the gene (e.g., the zinc-finger domains) are embedded within regions that are less densely packed with repetitive motifs; however, close analysis of the immediate upstream and downstream sequences of the above elements reveals that such regions are not refractory to the mutational processes producing simple-sequence DNA.
- Interspecific and intergeneric comparisons of the \textit{hunchback} region reveal extensive sharing of the same repetitive motifs, the precise extent of which depends on the specific pairwise comparison.
- Changes in BICOID-binding-site copy number, orientation, and spacing have involved a history of possible duplications and losses. Changes in binding-site sequence have taken place within diverse sites that predated the \textit{Drosophila/Musca} divergence time of approximately 100 Myr, given that individual orthologous sites of each species cluster together in a phylogenetic comparison.
- Changes in BICOID-binding-site sequences affect the strengths of binding of BICOID in in vitro assays.

From the above list of observations, we conclude that DNA slippage and slippage-like processes are the predominant mechanisms of turnover within and surrounding the \textit{hunchback} gene. Extensive sharing of specific short repetitive motifs, horizontally among different regions (both adjacent and nonadjacent) and vertically between orthologous regions between species, has been observed before in the so-called “expansion segments” of eukaryotic ribosomal RNA genes (Hancock and Dover 1988; Hancock 1995a). Cryptic simplicity (that is, regions with greater-than-expected numbers of a few motifs interspersed one with another) has been observed across eukaryotic intergenic, intronic, and exonic sequences, and in eubacterial sequences in many genera (Tautz, Trick, and Dover 1986; Hancock 1995b, 1996c). Notwithstanding such earlier surveys, we are surprised at the extent of slippage-generated cryptic sequence repetition in the \textit{hunchback} region and the extent of the sharing of specific motifs within and between species.

When the appropriate analytical tools are used on sequences, it is possible to reveal much hidden structure within sequences, particularly as is the case with the 5' and 3' flanking regions of \textit{hunchback}, in which sequences are generally unalignable between species. The unalignability of sequences is not necessarily a reflection of uncontrolled neutral divergence of functionless sequences. The extensive sharing of motifs might signify correlated, coevolutionary changes reflecting some important function having to do with higher-order DNA structure and nucleosome phasing that is common to the regions in question in the different species.

As suggested by Pearson and Sinden (1998), repeat-induced high-order structures might give rise to alternative DNA structures involved in replication, repair, and recombination. Such structures are believed to be inherently flexible, capable of forming hairpins, triplexes, quadruplexes, and slipped-strand structures (for references to literature, see review by Pearson and Sinden 1998).

The finding that exonic sequences are also cryptically simple but for motifs not generally shared by the 5' and 3' flanking regions argues somewhat against the idea that the slippage-like processes are biased regarding the kinds of short motifs that they generate. Furthermore, we know from other studies (Hancock 1995a, 1995b) that the base composition of a sequence is not wholly responsible for the composition of the slippage-generated motifs in that sequence. In particular, it is possible that selection plays a role in promoting or eliminating slippage-generated motifs in exonic sequences. Coding regions rich in CAG/CAA probably reflect selection at two levels: (1) for in-frame products and (2) for acceptable amino acids. The main change between species is CAG→CAA, which is neutral. Might there be
a selective advantage in having repeats in a gene in order to help maintain low divergence in surrounding functionally important sequences such as the zinc-finger and other conserved regions? Mammalian genes containing long CAN repeats show very low synonymous substitution rates, (unpublished data), as do the flanks of long microsatellites (Schlötterer, Amos, and Tautz 1991; unpublished data).

Our analysis of the BICOID-binding sites and their immediate 5′ and 3′ sequences shows that these regions are subject to turnover and point mutational processes that have altered the overall organization of the P2 promoter and the particular sequences in the “cores” of the sites. Some of these changes affect the strength of BICOID binding to the individual sites. The full range of functional consequences of these changes as measured by in vitro and in vivo tests using mixtures of binding-site “cores” and immediate surrounding sequences will be reported elsewhere. We have evidence that the binding affinities of sites can be changed by sequence alterations outside of the immediate “cores.” It could be that the different configuration of sites with different affinities within a species is critical for the function of the promoter as a whole unit.

Because of the complexity of the P2 promoter and the hidden structure of the 5′-flank, we do not know as yet how the system is maintained while at the same time allowing taxon-specific differences to accrue. For the moment, it looks as if the regulatory 5′ and 3′ elements are operating “on the edge of chaos”; however, we believe that there is a hidden order consequential on extensive correlated coevolutionary changes within the hunchback region and possibly between the cis-acting elements and the trans-acting regulatory proteins.

Acknowledgments

We thank the Wellcome Trust for a grant to G.A.D., the BBSRC for a studentship to P.J.S., and the UK Medical Research Council for providing support to J.M.H.

LITERATURE CITED


MARCY K. UYENOYAMA, reviewing editor

Accepted October 19, 1998