J Mol Evol (1999) 49:558–566

JOURNAL OF MOLECULAR EVOLUTION

A Comparison of Homologous Developmental Genes from *Drosophila* and *Tribolium* Reveals Major Differences in Length and Trinucleotide Repeat Content

Karl J. Schmid,* Diethard Tautz

Institut für Genetik, Universität zu Köln, Weyertal 121, 50931 Köln, Germany

Received: 13 January 1999 / Accepted: xxx

Abstract. The flour beetle *Tribolium castaneum* has become an important model organism for comparative studies of insect development. Many developmentally important genes have now been cloned from both *Tribolium* and *Drosophila* and their expression characteristics were studied. We analyze here the complete coding sequences of 17 homologous gene pairs from *D. melanogaster* and *T. castaneum*, most of which encode transcription factors. We find that the *Tribolium* genes are on average 30% shorter than their *Drosophila* homologues. This appears to be due largely to the almost-complete absence of trinucleotide repeats in the coding sequences of *Tribolium* as well as the generally lower degree of internal repetitiveness. Clusters of polar and other amino acids such as glutamine, proline, and serine, which are often considered to be important for transcriptional activation domains in *Drosophila*, are almost completely absent in *Tribolium*. Codon usage is generally less biased in *Tribolium*, although we find a similar tendency for the preference of G- or C-ending codons and a higher bias in conserved subregions of the proteins as in *Drosophila*. Most of the aminoacid substitutions in the DNA-binding domains of the transcription factors occur at residues that do not make a specific contact to DNA, suggesting that the recognition sequences are likely to be conserved between the two species.

Key words: *Drosophila melanogaster* — *Tribolium castaneum* — Protein length — Trinucleotide repeats — Homopolymeric cluster — Transcription factor — Effective number of codons — DNA-binding domain

Introduction

Many aspects of early insect embryogenesis are controlled by transcription factors which act as regulatory switches controlling development (St Johnston and Nüsslein-Volhard 1992; Pankratz and Jäckle 1993). Transcription factors usually have a modular structure, with separate domains being involved in DNA-binding, transcriptional activation, dimerization, and subcellular localization (Mitchell and Tjian 1989; Triezenberg 1995). Transcription factors are classified by their DNA-binding domains in HOM/Hox, zinc-finger, basic helix–loop–helix (bHLH), and other classes of proteins (Nelson 1995). The interaction with DNA has been thoroughly studied on a structural and molecular level using isolated DNA binding domains for X-ray crystallography or NMR studies. In contrast, less is known about the structure and function of the other domains in the protein, particularly the transcriptional activation domains which interact with the proteins of the transcription initiation complex. Currently, three main types of activation domains are known from functional studies: acidic, glutamine-rich, and proline-rich domains (reviewed by Mitchell and Tijan 1989; Triezenberg 1995).

**Current address:* Section of Genetics and Development, Cornell University, Ithaca NY 14853, USA
Correspondence to: Diethard Tautz; *e-mail:* tautz@uni-koeln.de

Sequence comparisons have shown for many developmental genes that different parts of the proteins diverge at different rates. For example, in transcription factors the DNA-binding domains are usually highly conserved, while other regions diverge rapidly (e.g., Atchley et al. 1994; Purugganan et al. 1995). Still, most studies focus on comparative aspects of gene expression patterns, and other aspects of the molecular evolution of the proteins are often not analyzed. Compared to metabolic enzymes or structural proteins, relatively little is therefore known about the molecular evolution of developmental genes, although these have recently received more attention (Purugganan 1998).

Tribolium has become a particularly interesting species for comparative studies of developmental evolution (e.g., Tautz and Sommer 1995; Brown and Denell 1996; Wolff et al. 1998; Maderspacher et al. 1998; Brown et al. 1999). The sequence database of developmental genes from this species is growing, in particular, for transcriptional regulators. Comparative analyses show a conservation of many expression patterns in both species despite significant differences in the modes of early development. In this study, we describe a comparative sequence analysis of 17 homologous genes whose complete coding sequences were determined from *Drosophila melanogaster* and *Tribolium castaneum*. The genes comprise 15 transcription factors regulating early embryonic development, 1 signaling molecule (*decapentaplegic*), and 1 metabolic enzyme (*Amylase*). The homologous gene pairs were compared with respect to sequence length and complexity, codon usage, and conservation of structural features that are thought to be important for their function.

Methods

Accession Numbers

We have analyzed published and unpublished sequences from *Drosophila melanogaster* and *Tribolium castaneum*. The GenBank accession numbers or sources for the *D. melanogaster* (DRO) and *Tribolium* (TRI) sequences are as follows: *Abdominal A* (DRO; X54453; TRI, AF017415), α-*Amylase* (DRO, L22716; TRI, TCU04271), *caudal* [DRO, M21069 and M21070; TRI, AJ005421; only the longer variant (cad-A; Schulz et al. 1998) was used], *decapentaplegic* (DRO, M30116; TRI, TCU63132), *Deformed* (DRO, X05136; TRI, TCU81038), *Distal-less* (DRO, S47947; TRI, A Beermann and D Tautz, unpublished), *Dorsal* (DRO, M23702; TRI, S Roth, unpublished), *empty spiracles* (DRO, X51653; TRI, B Hausdorf and D Tautz, unpublished), *engrailed* (DRO, M10017; TRI, S73255), *even-skipped* (DRO, M14767; TRI, TCU77974), *fushi-tarazu* (DRO, X00854; TRI, TCU14732), *hairy* (DRO, X15905; TRI, S Brown, unpublished), *hunchback* (DRO, Y00274; TRI, X91618), *orthodenticle* [DRO, X58983; TRI, AJ223627; only the longer variant (otd-1; Li et al. 1996) was used], *runt* (DRO, X55719; TRI, S Brown, unpublished), *tailless* (DRO, M34639; TRI, R Schröder and D Tautz, unpublished), and *zerknüllt* (DRO, X68346; TRI, X97819).

Sequence Analysis

The lengths of the complete open reading frames were obtained from database entries or by searching for the longest frame beginning with a methionine. Sequences were aligned with the CLUSTALW algorithm (Thompson et al. 1994), with some manual corrections. The relative simplicity factor RSF (Tautz et al. 1986) was calculated with the program SIMPLE34 (Hancock and Armstrong 1994). Codon usage and GC content were determined with the program CODONS (Lloyd and Sharp 1992). The effective number of codons (ENC) (Wright 1990) was used as a measure of codon usage bias. To compare codon usage patterns between species and regions, codon usage data were converted to relative synonymous codon usage (RSCU) values (Sharp et al. 1986). RSCU values lower than 1 indicate that a codon is avoided, and values higher than 1 that it is preferred.

The amino acid sequences were also scanned for features such as homopolymeric runs and distinct charge clusters. Stretches of identical amino acids were counted as distinct homopolymeric runs, if their length was at least five residues. Statistically significant clusters of polar and other (G, S, P) amino acids were identified with the SAPS program (Brendel et al. 1992), which employs an algorithm of Karlin and co-workers (1989).

Structural Analysis

Amino acid substitutions in the DNA-binding domains can affect the sequence specificity of DNA binding and may change the recognition sequence of a transcription factor. To analyze whether amino acid substitutions in the DNA-binding domain could affect binding specificity, substitutions that occurred in the DNA-binding domains of the transcription factors between *D. melanogaster* and *T. castaneum* were mapped onto structural models of these domains. For the homeodomains, the structural models of the *Antennapedia* (PDB code: 1AHD), *engrailed* (1ENH) and *fushi-tarazu* (1FTZ) homeodomains were used for comparison, and for the *hunchback* zinc-finger domains, the structures of *Zif268* (1ZAA) and *tramtrack* (1TTK). The bHLH domain of *hairy* was compared to *MyoD* (1MDY) and *Dorsal* to the *rel* domain of *NFkB* (1NFK). Finally, the receptor-binding domain of *decapentaplegic* was compared to the structure of *TGF*-β (1TFG). The locations and orientations of substituted residues were analyzed in sequence alignments and with the molecular viewer program RASMOL (Sayle and Milner-White 1995).

Results

Length Differences and Homopolymeric Runs

We find that the *Tribolium* proteins are on average almost 30% (145 amino acids) shorter than the *Drosophila* proteins (Table 1). This difference is highly significant (Wilcoxon signed rank test: $Z = -3.62$, $p < 0.001$). The protein that shows the smallest difference (4 amino acids) is the α-*Amylase* gene and the one with the largest difference is *orthodenticle* (300 amino acids). Much of the extra length of proteins in *Drosophila* appears to be due to amino acids repeats, because with the exception of *Abdominal A*, all *Drosophila* protein sequences contain a higher number of multiplets, that is, repeats of two or more identical amino acids ($Z = -3.26$, $p = 0.001$). Twelve of the 17 *Drosophila* proteins have one or more perfect homopolymeric run with a length of five or more

Table 1. Summary of sequence characteristics of gene pairs used in this study

		Length (aa)[a]		RSF[b]		ENC[c]	
Gene	Type	DRO	TRI	DRO	TRI	DRO	TRI
AbdominalA	HOX/*Hom*	330	284	2.11***[,d]	1.78***	51.87	49.62
Amylase	Enzyme	494	490	1.16*	1.16**	29.29	55.20
caudal	HOX/*Hom*	472	225	1.73***	1.22*	44.52	48.89
decapentaplegic	*TGF*-β	588	372	1.47***	1.15*	45.41	53.35
deformed	HOX/*Hom*	590	413	2.16***	1.38***	52.00	49.94
dorsal	*rel* domain	678	470	2.15***	1.15	51.62	58.01
Distal-less	HOX/*Hom*	327	313	1.07	1.31***	38.52	55.33
empty spiracles	HOX/*Hom*	494	282	1.43***	1.36***	42.94	38.38
engrailed	HOX/*Hom*	552	327	1.68***	1.26***	40.27	45.50
even-skipped	HOX/*Hom*	376	276	1.49***	1.41***	37.63	44.64
fushi-tarazu	HOX/*Hom*	414	322	1.39***	1.24**	39.36	59.26
hairy	bHLH	337	249	1.89***	1.44***	41.00	45.21
hunchback	Zinc finger	759	525	1.69***	1.32***	45.20	48.56
orthodenticle	HOX/*Hom*	671	371	2.60***	1.20***	48.81	51.75
runt	*runt* domain	509	369	1.45***	1.35***	36.88	42.26
tailless	NHR family	452	406	1.19***	1.23***	43.13	53.43
zerknüllt	HOX/*Hom*	353	242	1.27**	1.43**	48.45	59.38
Mean		482	342	1.60	1.32	42.83	50.04
p[e]		<0.001		0.008		0.003	

[a] Amino acids.
[b] Relative simplicity factor (Tautz et al. 1986).
[c] Effective number of codons (Wright 1990).
[d] Significance level of confidence interval (CI): *95% CI > 1.0; **99% CI > 1.0; ***99.7% CI > 1.0.
[e] Significance in a Wilcoxon signed-rank test.

residues, compared to only 2 of their *Tribolium* homologues (Table 2). Among the 38 homopolymeric runs, 15 are glutamine and 12 are alanine repeats.

Sequence Complexity

The degree of internal repetitivity of the DNA sequence can be quantitatively assessed by an algorithm identifying the nonrandom frequency distribution of short repeats within a narrow window (Tautz et al. 1986). The program counts the frequency of all possible tri- and tetranucleotide motifs in a given sequence and compares the frequencies to the average frequencies of random sequences with the same length and nucleotide composition. It then generates a relative simplicity factor (RSF) which has a value of 1.0 if the test sequence is not different from a random sequence. Using this algorithm it was shown that most eukaryotic genes contain a significantly higher than expected number of repeats in their coding regions (Tautz et al. 1986). This is also the case for the proteins studied here, but most of the *Drosophila* proteins have much higher relative simplicity factors than their *Tribolium* counterparts (Table 1). The average RSF for *Drosophila* is 1.61 and that for *Tribolium* is 1.32. Pairwise comparisons show that the difference is highly significant ($Z = -2.82$, $p = 0.005$).

Internal repeats for a given protein can also be visualized by self-similarity dot-plot comparisons. Figure 1 shows an example of such a comparison for the *orthodenticle* gene from both species. The multitude of internal repeats in *Drosophila* is evident, while these are virtually absent in *Tribolium*. It appears from this comparison that much of the extra length of the *orthodenticle* gene in *Drosophila* may be due to the internal repeats. In fact, across all gene pairs, length differences are positively correlated with differences in RSF values (Spearman rank correlation: $Z = 2.57$, $p = 0.01$).

Polar Clusters

Clusters of polar amino acids are hallmarks of transcription factors and other regulatory proteins (Brendel and Karlin 1989). They are virtually absent in metabolic enzymes and other housekeeping proteins (Karlin and Burge 1996). Polar clusters can be identified with an algorithm proposed by Karlin et al. (1989). First, the protein sequence is translated into a sequence of charges (positive, negative, and uncharged). From this sequence, the number of positively and negatively charged amino acids is summed up in a sliding window of 30 residues and compared to the number of polar residues that characterize a significant cluster given the length and total number of polar residues in the test sequence. Thirteen *Drosophila* and 10 *Tribolium* proteins contain at least one such cluster (Table 2). Many of these reside within the DNA-binding domains, which are usually positively charged (Brendel and Karlin 1989). In our sample of *Drosophila* transcription factors, we find seven proteins

Table 2. Amino acid repeats and clusters in homologous sequence pairs

	Multiplets		Homopolymeric runs		Polar clusters[a]		Other clusters	
	DRO	TRI	DRO	TRI	DRO	TRI	DRO	TRI
AbdominalA	25	33	Q_{17}, Q_6	A_5		±	Q	
Amylase	25	25	−	−	±			
caudal	50	16	H_5, N_{10}, N_7, R_{11}	−	+	±		
decapentaplegic	41	25	−	−	+, −	+		
deformed	57	35	$(HX)_5$, $(HX)_6$, G_5, Q_5, N_{11}, N_7	−	+	+, −	Q	
Distal-less	33	26	−	−		+		
dorsal	51	31	Q_5, Q_6, Q_7, Q_8, Q_{14}, N_8, A_5	−	±, ±		Q	
empty spiracles	51	23	Q_6, A_8	−	−	−		
engrailed	59	23	Q_{11}, A_{14}, A_6, A_7	−	−, ±		Q, S	
even-skipped	31	27	A_{11}	−	±	±		
fushi-tarazu	42	17	−	−	+			
hairy	32	23	Q_5, Q_6, Q_6, A_{10}	−	±	±		
hunchback	67	37	Q_5, H_5	N_5	−, −	−	Q, S	
orthodenticle	86	31	Q_5, A_5, A_5, A_5, A_7, G_{10}, $(GV)_6$, $(VR)_5$	−	+	+, ±	P	
runt	41	19	A_{11}, A_6		−	?		
tailless	38	37	−	−	?	?		P
zerknüllt	22	17	S_5	−				

[a] −, clusters of negative charge (E,D); +, positive charge (H,K,R); ±, mixed charge (any of the former).

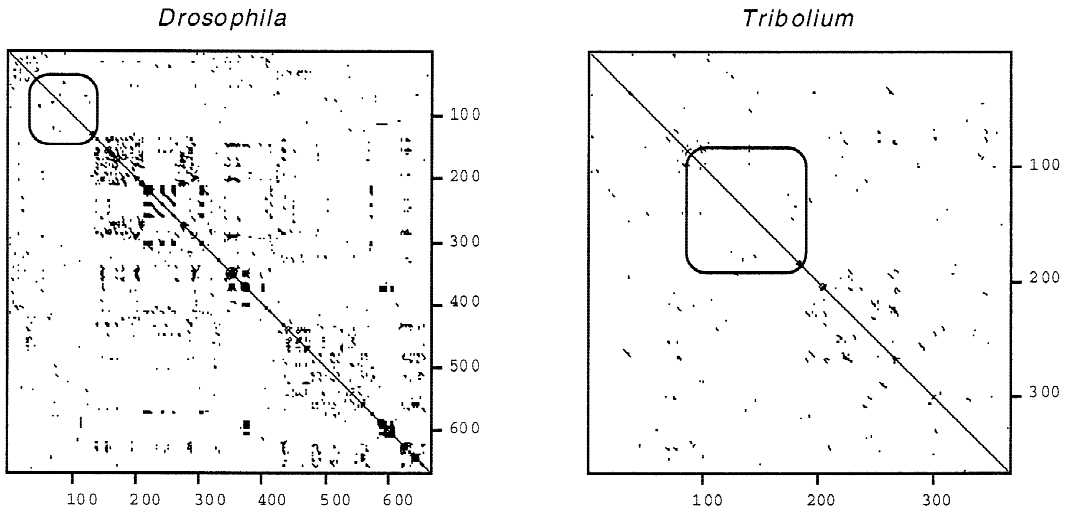

Fig. 1. Self-similarity dot-plot comparisons of the *orthodenticle* coding sequences from *D. melanogaster* and *T. castaneum.* A match stringency of three of five was chosen. Direct repeats within the sequence are represented by *squares of dots.* The region that is highly conserved between the two genes and which includes the homeodomain is *boxed.* Note the different scales for *Drosophila* and *Tribolium,* as the proteins are of different lengths.

with clusters outside of the DNA binding domain, but only three in *Tribolium,* suggesting that polar clusters outside DNA-binding domains are not strongly conserved sequence features of the transcription factors between *Drosophila* and *Tribolium.*

Codon Usage

The codon usage bias in *Tribolium* is clearly less pronounced than in *Drosophila.* Table 1 lists the ENC values for all proteins compared. On average, the ENCs are significantly higher than in *Drosophila,* indicating that *Tribolium* uses the less preferred codons more frequently. Table 3 lists the relative frequencies of the synonymous codons (RSCU) as they are found in our data set. The preferred codons for *Drosophila* match those that were derived from a much larger data set (Sharp and Lloyd 1993), indicating that the small data set analyzed here is representative. In *Tribolium,* most of the preferred codons are the same as in *Drosophila,* with notable differences only for the sixfold degenerate ones (Leu, Arg, and Ser). Still, the alternative codons preferred by *Tri-*

Table 3. Comparison of relative synonymous codon usage (RSCU) between *Drosophila melanogaster* and *Tribolium castaneum;* values for preferred codons are underlined

		Nonconserved regions		Conserved regions	
aa[a]	Codon	DRO	TRI	DRO	TRI
Leu	TTA	0.1	0.6	0.2	0.7
	TTG	0.8	1.4	0.5	0.9
	CTT	0.4	0.6	0.1	0.4
	CTC	1.1	1.5	1.2	1.8
	CTA	0.4	0.5	0.2	0.3
	CTG	3.2	1.5	3.7	1.9
Arg	CGT	0.5	0.5	0.9	0.6
	CGC	2.0	1.0	2.7	1.4
	CGA	0.8	0.9	0.8	0.7
	CGG	1.4	1.4	0.8	1.5
	AGA	0.5	0.5	0.3	0.9
	AGG	0.9	1.8	0.6	1.0
Val	GTT	0.8	0.8	0.5	0.6
	GTC	1.0	1.4	0.6	1.2
	GTA	0.3	0.2	0.5	0.3
	GTG	1.9	1.6	2.5	1.9
Thr	ACT	0.4	0.5	0.3	0.4
	ACC	1.6	1.6	2.2	1.8
	ACA	0.8	0.6	0.5	0.7
	ACG	1.1	1.4	1.0	1.1
Tyr	TAT	0.6	0.6	0.4	0.6
	TAC	1.4	1.4	1.6	1.4
His	CAT	0.8	0.5	0.5	0.3
	CAC	1.2	1.5	1.5	1.7
Gln	CAA	0.4	0.8	0.4	0.9
	CAG	1.6	1.2	1.6	1.1
Asn	AAT	0.8	0.7	0.5	0.7
	AAC	1.2	1.3	1.5	1.3
Lys	AAA	0.5	1.0	0.3	1.0
	AAG	1.5	1.0	1.7	1.0
Ser	TCT	0.4	0.3	0.2	0.2
	TCC	1.3	1.4	1.2	1.3
	TCA	0.3	0.5	0.1	0.4
	TCG	1.3	1.6	2.1	1.5
	AGT	0.7	1.2	0.4	1.0
	AGC	1.9	1.0	2.0	1.6
Gly	GGT	0.8	0.4	0.3	0.6
	GGC	1.9	1.6	2.3	1.6
	GGA	1.0	0.9	1.1	0.4
	GGG	0.3	1.1	0.3	1.3
Phe	TTT	0.5	0.7	0.4	0.6
	TTC	1.5	1.3	1.6	1.4
Pro	CCT	0.4	0.4	0.2	0.2
	CCC	1.6	1.6	1.8	1.7
	CCA	0.8	0.6	0.6	0.7
	CCG	1.3	1.4	1.5	1.4
Ala	GCT	0.5	0.6	0.2	0.4
	GCC	2.0	1.8	2.6	2.5
	GCA	0.7	0.4	0.3	0.4
	GCG	0.7	1.2	0.9	0.7
Asp	GAT	0.9	0.6	0.7	0.6
	GAC	1.1	1.4	1.3	1.4
Glu	GAA	0.4	0.8	0.4	0.8
	GAG	1.6	1.2	1.6	1.2
Cys	TGT	0.6	1.1	0.5	0.5
	TGC	1.4	0.9	1.5	1.5
Ile	ATT	0.7	1.1	0.7	0.6
	ATC	1.9	1.4	2.2	2.1
	ATA	0.4	0.4	0.1	0.3

[a] Amino acid.

bolium in these cases are the ones that end with a G or C, supporting a general bias for G- or C-ending codons in both species (Sharp and Lloyd 1993; Moriyama and Hartl 1993). Nonetheless, the RSCU values are lower for *Tribolium*, again indicating a generally lower preference. On the other hand, Akashi (1994) has shown that the highly conserved domains within *Drosophila* proteins show a stronger bias in codon usage, and this tendency is also evident in our data set, for both *Drosophila* and *Tribolium* (Table 3).

Substitutions in Conserved Domains

Amino acid substitutions in the conserved DNA- and ligand-binding domains could affect the specificity of DNA–protein or protein–protein interactions. In our sample, there are very different degrees of sequence conservation in these domains. For example, in the 60-amino acid-long homeodomain of the 10 HOM/Hox proteins, the numbers of substitutions range from 0 in *Abdominal A, Deformed,* and *Distal-less* to 20 in *zerknüllt.*

For most domains, a protein structure is available and therefore substitutions between *D. melanogaster* and *T. castaneum* can be mapped on their three-dimensional structure. A consistent pattern is observed among the DNA-binding domains. Nearly all substitutions affect residues that are oriented away from the DNA and exposed to the surface of the domain. In the DNA-binding domains of the HOM/Hox, *Dorsal*, and *hairy* proteins, none of the residues which make specific or nonspecific contacts to DNA were substituted between *Drosophila* and *Tribolium* (Fig. 2A). Only a few substitutions are observed in the hydrophobic core. In the basic helix–loop–helix domain (bHLH) of *hairy*, substitutions are found in the C-terminal regions of helix 1 and helix 2, both of which are involved in the dimerization of *hairy;* no substitutions occur in the DNA-contacting region of helix 1. The same pattern is observed in the *rel* domain of *Dorsal.* No substitutions are found in the recognition helix, but several sites in the dimerization domain are divergent between *Drosophila* and *Tribolium.* The data suggest that the DNA-binding specificities of these proteins are conserved. However, the exact sequence specificity of DNA-binding domains also depends frequently on cooperative interactions with other transcriptional regulators (e.g., Mann and Khan 1996; Jun and Desplan 1996). Such interactions may be altered if substitutions occur in the surface areas of DNA-binding or dimerization domains of regulatory proteins. Since a substantial number of substitutions is observed in these domains in our sample, we cannot exclude that differences in regulatory interactions have evolved between *Drosophila* and *Tribolium.*

The results for the zinc-finger domains of *hunchback* are equivocal (Fig. 2B). Structural and statistical analy-

A

```
                    1         2         3         4         5         6
          1234567890123456789012345678901234567890123456789012345678901234567890
contact     M MPMP                  P   P   P           PPP   SP SS PSP P
core                      C   C    C     C        CC CC C    C  CC
cad  Dm   KDKYRVVYTDFQRLELEKEYCTSRYITIRRKSELAQTLSLSERQVKIWFQNRRAKERTSN
cad  Tc   ..........H..V.....FYY.........A...NS.G..................KQV
ems  Dm   PKRIRTAFSPSQLLKLEHAFESNQYVVGAERKALAQNLNLSETQVKVWFQNRRTKHKRMQ
ems  Tc   ......................K.H........Q...S.S.T...................
en   Dm   EKRPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKIKKST
en   Tc   .........GA......H..A..............A.......................AS
eve  Dm   VRRYRTAFTRDQLGRLEKEFYKENYVSRPRRCELAAQLNLPESTIKVWFQNRRMKDKRQR
eve  Tc   I.........E..A......F.....................................
ftz  Dm   SKRTRQTYTRYQTLELEKEFHFNRYITRRRRIDIANALSLSERQIKIWFQNRRMKSKKDR
ftz  Tc   N......................K.L.....E..ES.R.T..............A...T
otd  Dm   QRRERTTFTRAQLDVLEALFGKTRYPDIFMREEVALKINLPESRVQVWFKNRRAKCRQQL
otd  Tc   ..............L..G..A..............V.......................
zen  Dm   LKRSRTAFTSVQLVELENEFKSNMYLYRTRRIEIAQRLSLCERQVKIWFQNRRMKFKKDI
zen  Tc   G..A...Y..A......R..HHGK..S.P...Q..EN.N.S...I..........H..EQ
```

B

```
                    1         2         3
          12345678901234567890123456789012
contact     P           P   PSPSS   SP       P
Zif268-1  ERPYACPVESCDRRFSRSDELTRHIRI--HTG
Zif268-2  QKPFQCRI--CMRNFSRSDHLTTHIRT--HTG
Zif268-3  EKPFACDI--CGRKFARSDERKRHTKI--HLR
ttk-1     EHTYRCKV--CSRVYTHLSNFCRHYVTS-HKR
ttk-2     VKVYPCPF--CFKEFTRKDNMTAHVKII-HKI
hb-1 Dm   MKNYKCKT--CGVVAITKVDFWAHTRT--HMK
hb-1 Tc   I.TF...Q--.DF.....LEQ.N.SKV--.IR
hb-2 Dm   DKILQCAK--CPFVTEFKHHLEYHIRK--HKN
hb-2 Tc   ..R.T.P.--...IT.Y.......L.N--.AG
hb-3 Dm   QKPFQCDK--CSYTCVNKSMLNSHRKS--HSS
hb-3 Tc   S.....N.--.D...........M..--..N
hb-4 Dm   VYQYRCAD--CDYATKYCHSFKLHLRKYGHKP
hb-4 Tc   ..R.S.R.--.S........L.I...R...T.
hb-5 Dm   PAIYECKY--CDIYFKDAVLYTIHMGY--HSC
hb-5 Tc   EEGNS.Q.--.N.A.G.............--.GF
hb-6 Dm   DDVFKCNM--CGEKCDGPVGLFVHMARNAHS
hb-6 Tc   HNP.T...--..VE.SDK.SF.LHI..VS..
```

Fig. 2. Sequence alignment of homeodomain and zinc-finger domains to demonstrate protein–DNA contacts. The structural roles of amino acids was taken from the PDB files and the literature and are designated as follows: C, contributes to the hydrophic core; P, contacts the sugar–phosphate backbone; S, forms specific contacts in the major groove; M, forms unspecific contacts in the minor groove. The amino acids contacting the DNA are in *boldface.* Note that not all contacts are made by every domain, but the figure helps to estimate which amino acids may or may not be involved in the contacts. Dm, *D. melanogaster;* Tc, *T. castaneum.* **A** Homeodomains; *Antennapedia* is used as a reference. **B** Zinc fingers; *Zif268* and *tramtrack* are used as a reference.

ses originally identified three residues making specific contacts with DNA bases (Pavletich and Pabo 1991; Jacobs 1992). However, further studies identified additional residues that make specific contacts to DNA (Pavletich and Pabo 1993). Some of the numerous substitutions in the six zinc fingers of *hunchback* (Fig. 2B) occur at positions that were shown to make specific contacts with DNA and thus could change the recognition sequence between the two species.

Substitutions in the *TGF*-β-like domain of the secreted *decapentaplegic* protein occur only at exposed residues. Since the receptor-binding residues are not exactly known (Daopin et al. 1992), it remains unknown whether the observed substitutions change ligand–receptor interactions.

Discussion

The most conspicuous difference between the *Drosophila* and the *Tribolium* genes studied here is that the *Tribolium* genes are markedly shorter and less repetitive than their *Drosophila* homologues. Furthermore, we observe an almost-complete absence of homopolymeric stretches of single amino acids in *Tribolium,* which are a prominent feature of many *Drosophila* developmental

genes (Karlin and Burge 1996). Are these differences caused by general genomic differences (i.e., a high rate of replication slippage that may lead to a mutational pressure toward longer genes in *Drosophila*) or do they reflect functional differences in the proteins between the two species? Also, one may ask whether it is the *Drosophila* or the *Tribolium* genome that is more unusual in this respect.

Hancock (1996) has suggested that there is a relationship between the repetitiveness of genomic sequences and the genome size. In this analysis, it seemed that it is the *Drosophila* genome which is unusual, as it shows a high general repetitiveness, despite its relatively small size. Since the genome sizes of *Drosophila* and *Tribolium* are similar [$1.8 * 10^8$ and $2.1 * 10^8$ nucleotides (Alvarez-Fuster et al. 1991)], one would conclude that a lower degree of repetitiveness of genes in *Tribolium* is more in line with the expectations. There are currently only two sequences available from long noncoding regions in *Tribolium* that can be analyzed in a similar way. One is the large intron in the *decapentaplegic* gene, which has an RSF of 1.36 in *Drosophila* and 1.12 in *Tribolium.* The other is the upstream region from *hunchback,* with 1.81 in *Drosophila* and 1.23 in *Tribolium.* The values from two as yet unpublished upstream regions from *hairy* (*h*) and *tailless* (*tll*) are as follows: *Drosophila h,* 1.83; *Tribolium h,* 1.40; *Drosophila tll,* 1.32; and *Tribolium tll,* 1.57. Thus, there is a slight, although not consistent tendency for *Drosophila* regions having the higher RSF values. Accordingly, there might be a different mutational pressure in the *Drosophila* genome that results in a higher likelihood of generating such internal repeats. However, if a strong mutational pressure increased sequence repetitivity in coding and noncoding regions in the same way, larger differences in genome size would be expected. Also, only 4 of the 39 homopolymeric runs in the *Drosophila* genes are encoded by perfect trinucleotide repeats of single codons. This suggests either that replication slippage is not the exclusive mechanism for generating these repeats or that the repeats were retained after their generation because they had some functional importance and were subsequently saturated with synonymous substitutions. Finally, mostly genes involved in developmental control or neurogenesis contain repetitive sequences in *Drosophila,* while metabolic enzymes and other housekeeping proteins do not (Karlin and Burge 1996). This is supported by the comparison of the *Amylase* homologue from *Drosophila* and *Tribolium,* which is the only nondevelopmental protein in our sample (Tables 1 and 2). This protein does not contain any homopolymeric repeats or other clusters, and its length and other sequence characteristics are very similar between the two species. It should also be noted that the homeodomain proteins in our sample show rather different degrees of repetitiveness. Among the *Drosophila* genes, *Distal-less* (RSF: 1.07) is clearly less repetitive than *orthodenticle* (RSF: 2.60) (Fig. 1), and accordingly, length differences between *Drosophila* and *Tribolium* are smaller in the former gene (14 vs 300 amino acids; Table 1). Taken together, these observations do not support the hypothesis of a general (e.g., mutational) difference between the two genomes.

The alternative to mutational differences in the genomes are differences in functional constraints. Little is known about a functional role of homopolymeric repeats in transcription factors so far. One suggested role is that homopolymeric repeats may function in transcriptional activation because moderate-length homopeptides of glutamine and proline were found to stimulate transcriptional activation of yeast genes in vivo and in vitro (Gerber et al. 1994). It was also shown that the glutamine-rich domains of *Drosophila Antennapedia* and human *Sp1* transcription factors interact in vitro with the TATA-binding protein (TBP) (Emili et al. 1994) and TBP-associated proteins (TAFs) (Gill et al. 1994). Another possibility is that homopolymeric repeats serve as spacers of small conserved charge clusters or other functional motifs that interact with DNA or other proteins. These spacers may act as flexible hinges facilitating three-dimensional interactions among transcription factors, which are bound to distant regulatory elements. However, it is difficult to imagine how such generic functions could be so drastically different between *Drosophila* and *Tribolium.*

Arguments against a specific function of such repeats stem from observations that the length of homologous repeats can vary significantly within and between *Drosophila* species (Michalakis and Veuille 1996; Treier et al. 1989; Tautz and Nigro 1998) and that, in some cases, insertion mutations change glutamine (encoded by CAG) to alanine (GCA) repeats. Glutamine and alanine form the largest number of homopolymers in our sample and also in the survey by Karlin and Burge (1996). On the other hand, such repeats are not likely to be entirely neutral either, because expansions of glutamine repeats can become unstable and cause neurodegenerative disorders in humans (Bates and Lehrach 1994). If the repeats were functionless, all replacement substitutions in these regions should be neutral. However, several homopolymers in this study are saturated with silent mutations, suggesting that they are relatively old and subject to purifying selection.

We also find a relative lack of clusters of polar amino acids, glutamine, serine, and threonine in the putative transactivation domains of *Tribolium* transcription factors, although they are supposed to play a role in transcription. The small number of clear clusters in *Tribolium* genes indicates that they are not essential for stimulating transcriptional activation. Activation may be achieved by other, more subtle and as yet unknown motifs. For example, bulky hydrophobic amino residues in the glutamine-rich domain of human *Sp1* and polar re-

gions of yeast *GAL4* transcription factors must also be present to achieve transcriptional activation, presumably by forming an amphipathic α-helix with polar amino acids (Ruden 1992; Gill et al. 1994). Further functional and structural analyses will be necessary to resolve this question.

Comparative expression analysis suggests that the genes studied here are likely to have similar developmental functions in the two species (reviewed by Tautz and Sommer 1995; Brown and Denell, 1996). This is also supported by our mapping of amino acid substitutions on the structural model of DNA-binding domains, which suggest that there has been no major change in DNA binding specificity. While the postulated activation domains seem to exhibit strong differences that may affect the specificity and type of protein–protein interactions in both species, the analysis of substitutions in the DNA-binding domains shows that most of them do not affect the specificity of DNA binding. Most residues with substitutions are oriented away from the DNA molecule and are exposed toward the surface of the protein. We therefore expect that the protein–DNA interactions with cis-regulatory elements of target genes are conserved between *Drosophila* and *Tribolium.* In fact, two *Tribolium* genes, *caudal* and *hunchback*, that were transferred into *Drosophila* were found to be faithfully regulated by the *Drosophila* transcriptional and translational machinery (Wolff et al. 1998). The only difference that was found concerned a whole promoter that appears to have been lost in the *hunchback* upstream region in the evolutionary line toward *Drosophila* (Wolff et al. 1998). Brown et al. (1999) have recently undertaken a direct test of the protein function of *Tribolium Deformed (Dfd)* in *Drosophila* and could show that it activates known target genes of *Dfd* in *Drosophila* in the same way as the endogeneous *Dfd.* Moreover, the *Tribolium* protein did rescue a *Dfd* mutant phenotype to the same extent as the *Drosophila* protein. However, subtle differences in regulatory functions would not have been noted in all these experiments. Still, these results underpin our inference that the most important sites in the proteins are conserved.

Both the *Drosophila* and the *Tribolium* genes show an unequal usage of synonymous codons, but the codon usage is more biased in *Drosophila.* A simple reason for this could be different mutational patterns, for example, a higher tendency to incorporate A or T nucleotides. The *Tribolium* genes do indeed show a tendency for a lower G/C content than the *Drosophila* homologues and noncoding regions of *Tribolium* tend to be fairly A/T rich. Another potential reason for the more pronounced codon usage in *Drosophila* could be selection for higher translational accuracy (Akashi 1994). Since embryogenesis is significantly shorter in *Drosophila* (24 h) than in *Tribolium* (7 days), the need for rapid and accurate protein synthesis may be more important, thus enhancing selection on optimal codon usage. A further explanation could be a lower effective population size in *Tribolium castaneum.* Selection coefficients on silent sites are very low (Shields et al. 1988; Akashi 1995) and natural selection may not be efficient enough in *Tribolium* to overcome the effects of random drift, causing a less biased usage of synonymous codons. However, this possibility remains speculative, because little is known about the population size and structure of *T. castaneum.*

Conclusion

Our data show a major difference in sequence length and complexity of regulatory genes between *Drosophila* and *Tribolium.* How far this reflects functional differences has yet to be analyzed using appropriate transgenic or biochemical approaches. Also, the evolutionary history of these differences has to be investigated further, in particular, the question at which point these differences originated. Comparisons of the *hunchback* gene from different species (Hancock et al. 1999) suggest that this has happened in the line to the higher diptera, but additional comparisons with other genes from multiple taxa will be necessary to verify this inference.

Acknowledgments. We are grateful to A. Beermann, S. Brown, B. Hausdorf, S. Roth, and R. Schröder for providing unpublished sequence data for this analysis.

References

Akashi H (1994) Synonymous codon usage in *Drosophila melanogaster:* Natural selection and translational accuracy. Genetics 136:927–935

Akashi H (1995) Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in *Drosophila* DNA. Genetics 139:1067–1076

Alvarez-Fuster A, Juan C, Petitpierre E (1991) Genome size in *Tribolium* flour-beetles: Inter- and intraspecific variation. Genet Res Cambr 58:1–5

Atchley WR, Fitch WM, Bonner-Fraser M (1994) Molecular evolution of the *MyoD* family of transcription factors. Proc Natl Acad Sci USA 91:11522–11526

Bates G, Lehrach H (1994) Trinucleotide repeat expansions and human genetic disease. BioEssays 16:277–284

Brendel V, Karlin S (1989) Association of charge clusters with functional domains of cellular transcription factors. Proc Natl Acad Sci USA 86:5698–5702

Brendel V, Bucher P, Nourbakshs I, Blaisdell E, Karlin S (1992) Methods and algorithms for statistical analysis of protein sequences. Proc Natl Acad Sci USA 89:2002–2006

Brown SJ, Denell RE (1996) Segmentation and dorsoventral patterning in *Tribolium.* Semin Cell Dev Biol 7:553–560

Brown S, Holtzman S, Kaufman T, Denell R (1999) Characterization of the *Tribolium Deformed* ortholog and its ability to directly regulate *Deformed* target genes in the rescue of a *Drosophila Deformed* null mutant. Dev Genes Evol 209:389–398

Daopin S, Piez K, Ogawa Y, Davies D (1992) Crystal structure of transforming growth factor-β 2: An unusual fold for the superfamily. Science 257:369–373

Emili A, Greenblatt J, Ingles CJ (1994) Species-specific interaction of

the glutamine-rich activation domains of *Sp1* with the TATA box-binding protein. Mol Cell Biol 14:1582–1593

Gerber HP, Seipel K, Georgiev O, et al. (1994) Transcriptional activation modulated by homopolymeric glutamine and proline stretches. Science 263:808–811

Gill G, Pascal E, Tseng ZH, Tjian R (1994) A glutamine-rich hydrophobic patch in transcription factor *Sp1* contacts the $dTAF_{II}110$ component of the *Drosophila* TFIID complex and mediates transcriptional activation. Proc Natl Acad Sci USA 91:192–196

Hancock JM (1996) Simple sequences and the expanding genome. Bioessays 18:421–425

Hancock JM, Armstrong JS (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. Comp Appl Biol Sci 10:67–70

Hancock JM, Shaw PJ, Bonneton F, Dover GA (1999) High sequence turnover in the regulatory regions of the development gene *hunchback* in insects. Mol Biol Evol 16:253–265

Jacobs GH (1992) Determination of the base-recognition of zinc-fingers from sequence-analysis. EMBO J 11:4507–4517

Jun S, Desplan C (1996) Cooperative interactions between paired domain and homeodomain. Development 122:2639–2650

Karlin S, Burge C (1996) Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. Proc Natl Acad Sci USA 93:1560–1565

Karlin S, Blaisdell BE, Mocarski ES, Brendel V (1989) A method to identify distinctive charge configurations in proteins, with application to human herpesvirus polypeptides. J Mol Biol 205:165–177

Li Y, Brown S, Denell R, Hausdorf B, Tautz D, Finkelstein R (1996) Two *orthodenticle* related genes in the short-germ beetle *Tribolium castaneum.* Dev Genes Evol 206:35–45

Lloyd AT, Sharp PM (1992) CODONS: A microcomputer program for codon usage analysis. J Hered 83:239–240

Maderspacher F, Bucher G, Klingler M (1998) Pair-rule and gap gene mutants in the flour beetle *Tribolium* castaneum. Dev Genes Evol 208:558–568

Mann R, Khan SK (1996) Extra specificity from *extradenticle:* The partnership between HOX and PBX/EXD homeodomain proteins. Trends Genet 12:258–262

Michalakis Y, Veuille M (1996) Length variation of CAG/CAA trinucleotide repeats in natural populations of *Drosophila melanogaster* and its relation to the recombination rate. Genetics 143:1713–1725

Mitchell PJ, Tjian R (1989) Transcriptional regulation im mammalian cells by sequence-specific DNA-binding proteins. Science 245: 371–378

Moriyama EN, Hartl DL (1993) Codon usage bias and base composition of nuclear genes in *Drosophila.* Genetics 134:847–858

Nelson H (1995) Structure and function of DNA-binding proteins. Curr Opin Gen Dev 5:180–189

Pankratz M, Jäckle H (1993) Blastoderm Segmentation. In: Bate M, Martinez-Arias A (eds) *Development of Drosophila malanogaster*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, pp 467–516

Pavletich NP, Pabo CO (1991) Zinc finger-DNA recognition: Crystal structure of a *Zif268*-DNA complex at 2.1Å. Science 252:809–817

Pavletich NP, Pabo CO (1993) Crystal-structure of a 5-finger GLI-DNA complex crystal—New perspectives on zinc fingers. Science 261:1701–1707

Purugganan MD (1998) The molecular evolution of development. BioEssays 20:700–711

Purugganan MD, Rounsley SD, Schmidt RJ, Yanofsky MF (1995) Molecular evolution of flower development: Diversification of the plant MADS-box regulatory gene family. Genetics 140:345–356

Ruden DM (1992) Activating regions of yeast transcription factors must have both acidic and hydrophobic amino acids. Chromosoma 101:342–348

Sayle RA, Milner-White EJ (1995) RASMOL: Biomolecular graphics for all. Trends Biochem Sci 20:374

Schulz C, Schröder R, Hausdorf B, Wolff C, Tautz D (1998) A *caudal* homologue in the short germ band beetle *Tribolium* shows similarities to both, the *Drosophila* and the vertebrate *caudal* expression patterns. Dev Genes Evol 208:283–289

Sharp PM, Lloyd AT (1993) Codon Usage. In: Maroni G (ed) An atlas of *Drosophila* genes: Sequences and molecular features. Oxford University Press, New York, pp 378–397

Sharp PM, Tuohy TMF, Mosurski K (1986) Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res 14:5125–5143

Shields DC, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in *Drosophila* are not neutral: Evidence of selection among synonymous codons. Mol Biol Evol 5:704–716

St Johnston D, Nüsslein-Volhard C (1992) The origin of pattern and polarity in the *Drosophila* embryo. Cell 68:201–219

Tautz D, Nigro L (1998) Microevolutionary divergence pattern of the segmentation gene *hunchback* in *Drosophila.* Mol Biol 15:1403–1411

Tautz D, Sommer R (1995) Evolution of segmentation genes in insects. Trends Genet 11:23–37

Tautz D, Trick M, Dover GA (1986) Cryptic simplicity in DNA is a widespread source of genetic variation. Nature 322:652–656

Thompson JD, Higgins DG, Gibson TJ (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

Treier M, Pfeifle C, Tautz D (1989) Comparison of the gap segmentation gene *hunchback* between *Drosophila melanogaster* and *Drosophila virilis* reveals novel modes of evolutionary change. EMBO J 8:1517–1525

Triezenberg SJ (1995) Structure and function of transcriptional activation domains. Curr Opin Genet Dev 5:190–196

Wright F (1990) The "effective number of codons" used in a gene. Gene 87:23–29

Wolff C, Schröder R, Schulz C, Tautz D, 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