Interspecific comparison of the *period* gene of *Drosophila* reveals large blocks of non-conserved coding DNA

Hildur V.Colot, Jeffrey C.Hall and Michael Rosbash

Department of Biology, Brandeis University, Waltham, MA 02254, USA

Communicated by M.Heisenberg

We have cloned and sequenced the coding region of the *period* (*per*) gene from *Drosophila pseudoobscura* and *D.virilis*. A comparison with that of *D.melanogaster* reveals that the conceptual translation products consist of interspersed blocks of conserved and non-conserved amino acid sequence. The non-conserved portion, comprising ~33% of the protein sequence, includes the perfect Thr–Gly repeat of *D.melanogaster*, which is absent from the *D.pseudoobscura* and *D.virilis* proteins. Based on these observations and cross-species transformation experiments, we suggest that the interspecific variability in the *per* primary amino acid sequence contributes to the control of species-specific behaviors. **Key words:** *Drosophila pseudoobscura*/*Drosophila virilis*/biological rhythms/sequence comparisons/evolution

Introduction

The *period* (*per*) gene of *Drosophila melanogaster* influences the fly's circadian and ultradian rhythms; in the absence of the gene or its product, the fly does not manifest overt rhythms, whereas flies carrying mutant *per* genes display aberrant rhythms. The *per* gene has been sequenced, the amino acid sequence of the protein deduced, and the location of point mutations defined (for review, see Hall and Rosbash, 1988). Also, a considerable body of data now exists that describes the complicated pattern of *per* expression in both embryos and adults (James *et al.*, 1986; Liu *et al.*, 1988; Siwicki *et al.*, 1988). However, the biochemical functions of the *per* product(s) (Citri *et al.*, 1987) remain unknown.

Recent work from this laboratory presented evidence that regions of the *per* protein might contribute differentially to circadian rhythm phenotypes on the one hand and to ultradian rhythm (male courtship song) phenotypes on the other (Yu *et al.*, 1987a). These observations, coupled with the large size and an unusual primary sequence feature [a perfect repeat of 20 Thr–Gly (TG) pairs] of the predicted *per* protein, suggested that the *per* protein might consist of several functional domains. If only some of these domains contribute to the relatively conserved circadian rhythm phenotypes, then different regions of the *per* protein might manifest substantially different rates of divergence in interspecific comparisons.

To test this possibility, we have cloned and sequenced the coding region of the *per* gene from *D.pseudoobscura* and *D.virilis* and compared the data with those previously obtained for the *D.melanogaster* gene. The results show that

different regions of the protein indeed manifest vastly different rates of change. In light of these observations, we make speculations concerning the functional importance of rapidly evolving protein-coding subregions of proteins.

Results and discussion

Gene organization

Initial experiments aimed at detecting the *per* gene in *D.pseudoobscura* and *D.virilis* involved Southern blotting of genomic DNA digested with *Eco*RI and *Bam*HI. The probes used were two fragments from the *D.melanogaster per* gene: a 0.5 kb *Nco*I–*Sst*I (NS) fragment encompassing the site of the *per^S* mutation, and the adjacent 1.5 kb *Nco*I–*Nco*I fragment comprising exons 2–4 and the included introns (Yu *et al.*, 1987b; Baylies *et al.*, 1987; Citri *et al.*, 1987). Both fragments hybridized to unique restriction fragments from both *D.pseudoobscura* and *D.virilis* under low stringency conditions (see Materials and methods), although the 0.5 kb NS fragment gave a stronger signal (data not shown).

In additional digests of *D.pseudoobscura* and *D.virilis* genomic DNA with *Cfo*I, small unique fragments hybridized to the 0.5 kb NS fragment. These homologous fragments were cloned (see Materials and methods) for preliminary sequence analysis in order to ascertain that they were indeed derived from *bona fide per* genes. From this initial sequence analysis it was clear both that there is substantial homology among the putative *per* genes from the three species and that the *D.pseudoobscura* gene includes at least one region of remarkable divergence (see below). Based on these encouraging, preliminary results, we proceeded to clone and sequence the *per* genes from *D.pseudoobscura* and *D.virilis*.

The small subclones of the *D.pseudoobscura* and *D.virilis*

Fig. 1. Intron/exon structure of the *per* genes of *D.virilis*, *D.pseudoobscura* and *D.melanogaster*. Only the coding regions of each gene are depicted, and the introns (V) and exons (—) are drawn approximately to scale, such that each side of a V represents half the length of an intron. v: *D.virilis*; m: *D.melanogaster*; p: *D.pseudoobscura*. The non-coding 5' exon and the adjacent large intron in the *D.melanogaster* gene are not shown. The *D.virilis* introns (in the 5' to 3' direction) are 70, 461, 475, 72, 66 and 484 nt long; those of *D.pseudoobscura* are 62, 64, 532, 124, 62 and 197 nt long; and the *D.melanogaster* introns shown (cf. Citri *et al.*, 1987) are all between 57 and 70 nt long. The arrowheads indicate that non-coding portions of the genes extend in both directions.

Table I. Comparison of 'reduced' *per* genes

	Number of codons	Amino acid substitutions (%)	Synonymous substitutions (%)
<i>D.melanogaster</i> versus <i>D.pseudoobscura</i>	836	16.4	36.3
<i>D.melanogaster</i> versus <i>D.virilis</i>	812	19.1	38.2
<i>D.virilis</i> versus <i>D.pseudoobscura</i>	798	23.6	38.7

'Reduced' nucleic acid sequence files were created separately for each pair of proteins (see text) and compared by the SYNSUB program of Lewontin (1989). The number of codons in the complete putative proteins before 'reduction' were: *D.melanogaster*, 1218; *D.pseudoobscura*, 1241; and *D.virilis*, 1087. The values for synonymous substitutions reflect raw data, uncorrected for multiple hits, etc.

genes (see above) were used to screen the corresponding genomic libraries for clones of the entire genes by standard techniques (see Materials and methods). Only the coding regions were of interest in this study; sequencing was therefore not continued much beyond the putative initiation and termination codons, as determined by homology to the *D.melanogaster per* gene. A total of 5062 nt were sequenced for *D.pseudoobscura* and 4982 for *D.virilis*. The coding regions contain 4763 and 4893 nt, respectively, including the initiation and termination codons and the introns as shown in Figure 1. Our previous results predicted a *D.melanogaster per* primary translation product of 1218 amino acids; the data presented here predict *per* proteins of 1241 and 1087 amino acids in *D.pseudoobscura* and *D.virilis*, respectively. The inferred mol. wts, in daltons, of the putative proteins are 127 000 (*D.melanogaster*), 132 000 (*D.pseudoobscura*) and 118 000 (*D.virilis*).

Figure 1 shows the intron/exon maps of the three genes. The intron/exon junctions for the *D.melanogaster per* gene are known from cDNA clone analysis and RNase protection experiments (Jackson *et al.*, 1986; Citri *et al.*, 1987); the junctions in the *D.pseudoobscura* and *D.virilis* genes were deduced by comparison with the *D.melanogaster* sequence. In the 5' half of the coding portion of *D.melanogaster* gene, the three introns fall within highly conserved regions (see below) and are all within codons; thus their positions in the other two genes were predictable and, in each case, proper 5' and 3' splice junction sequences were found at exactly the predicted position. However, the 3' portions of the genes are much more divergent (see below) and, in *D.melanogaster*, the three introns lie between codons that are not always conserved. Consequently, the corresponding junctions in *D.pseudoobscura* and *D.virilis* were deduced from a combination of factors, including the position and frame of possible splice junction consensus sequences, the comparisons of adjacent coding sequences, and the relative 'fit' of candidate splice junctions with consensus splice junction sequences. In no case was the presence of an intron in doubt at approximately the predicted position, i.e. in both the *D.pseudoobscura* and *D.virilis* genes, there was always a stop codon (and usually a frame shift) between conserved coding regions located in two putative, adjacent exons. Also, selection of alternate splice junctions resulted in even less conservation of the *per* coding sequence (see below).

Remarkably, alignment of the sequences of the predicted *per* proteins from *D.melanogaster*, *D.pseudoobscura* and *D.virilis* is possible for only 60–70% of the protein. Moreover, the relatively conserved portions are interspersed with blocks of much less conserved sequences. It is unlikely that these are other than the true *per* genes of *D.pseudoobscura* and *D.virilis*, because of (i) the single-copy nature of the genes (see above and Materials and methods),

(ii) the high degree of conservation in major portions of the predicted proteins, (iii) the lack of stop codons in the alignable coding regions, (iv) the fact that even the non-conserved regions show certain interspecific similarities (see below), and (v) the fact that the regions of recognizable homology span the entire protein and include the N terminus and C terminus.

Conserved coding sequences

The conserved regions (c1–c6, Figure 3) are defined somewhat arbitrarily (see below) and include stretches in which any two of the three genes analyzed here are clearly homologous. It is readily apparent that there is a high degree of conservation at the amino acid level, especially in the N-terminal (c1) and C-terminal (c6) regions and in the longest conserved region (c2). The c2 region includes those sites to which the *per^{L1}* and *per^S* mutations have previously been mapped (Baylies *et al.*, 1987; Yu *et al.*, 1987b). In addition, it encompasses a region known to have homology to a portion of the *single-minded* gene (Crews *et al.*, 1988); this region includes two 51 amino acid direct repeats whose repeated features are conserved among the three species studied here, although a six amino acid stretch is notably different in the first copy of the *D.virilis* repeat (Figure 3). The serine at position 589 (which is replaced by an asparagine in the *per^S* mutation) is conserved in *D.pseudoobscura* and *D.virilis* and is included in one of the longest stretches of perfect conservation in the entire protein. The valine that is replaced by an aspartic acid in *per^{L1}* at position 243 is also conserved.

Due to the patchy nature and unusual extent of the divergence observed among these genes, it is difficult to compare our results with those obtained in other pair-wise comparisons of *Drosophila* genes. In order to approach this issue in a semi-quantitative way, we constructed separate pairs of 'reduced' nucleic acid sequence files for each pair of proteins by removing all but those regions in which the sequences could be clearly aligned by eye. The resulting files therefore included only comparable regions of each protein. The sequences of these 'reduced' proteins were then compared by the SYNSUB program of Lewontin (1989) to determine the frequency of both synonymous and non-synonymous nucleotide substitutions (see Table I).

In general, the extent of divergence at the amino acid level between pairs of 'reduced' *per* genes is higher than that observed for most of the other genes that have been compared among *D.melanogaster*, *D.virilis* and *D.pseudoobscura*. Published values are typically 3–15% (Blackman and Meselson, 1986; Henikoff and Eghedarzadeh, 1987; Schaeffer and Aquadro, 1987; Aguade, 1988; Riley, 1989). Our values obtained from pairwise comparisons range from 16% for *D.melanogaster* versus *D.pseudoobscura* to 24%

for *D.pseudoobscura* versus *D.virilis* (Table I).

Of the synonymous positions, 36–39% have accumulated substitutions (note that these raw data have not been corrected for multiple hits, etc.). These values are within the range of those obtained in comparisons of several other genes from these species (e.g. Blackman and Meselson, 1986; Henikoff and Eghtedarzadeh, 1987; Schaeffer and Aquadro, 1987; Wilde and Akam, 1987; Aguade, 1988; Riley, 1989).

Non-conserved coding regions

It is readily apparent that a substantial fraction of the *per* gene ($\geq 30\%$) has undergone much higher rates of evolution than other regions. In addition, the fraction of rapidly evolving coding DNA is, to our knowledge, higher than in any pair of *Drosophila* genes compared to date. Only the *engrailed* (*en*) and *Ultrabithorax* (*Ubx*) genes show a similar phenomenon, and they do so to a much lesser extent (Kassis *et al.*, 1986; Wilde and Akam, 1987). In *per* (as in *en* and *Ubx*), the non-conserved regions are clustered, such that, aside from a long stretch of from 116 to 153 amino acids near the N termini of the *per* proteins, most of the divergence has occurred in the C-terminal halves. Also, the strict divisions we have made between conserved and non-conserved regions, especially in the C-terminal half of the protein, are necessarily somewhat arbitrary. Although the non-conserved stretches cannot be clearly aligned, they do have some features in common as seen upon inspection by eye, by nucleic acid homology searches, and on hydropathy plots of the primary amino acid sequences (Pustell and Kafatos, 1986). We will discuss each region, using the designations (n1–n5) defined in Figures 2 and 3.

The large, most N-terminal non-conserved region, n1, is notable for its appearance within an otherwise highly conserved portion of the protein, regions c1 and c2. The first 60% of each n1 region is hydrophilic and is followed by a (short) hydrophobic stretch leading into another hydrophilic region which continues into the adjacent conserved region (c2). Similarly, near the end of c2 the protein sequence undergoes a transition from hydrophilic to hydrophobic through a short stretch that is conserved between the *D.pseudoobscura* (Figure 3, 605–621) and *D.virilis* (625–638) proteins but quite different in the *D.melanogaster* protein (643–660).

The non-conserved region n2 is the most outstanding in both the nature of its sequences and the extent of the divergence observed. Notably, the Thr–Gly (TG) repeat (see below) is not conserved among the two other distantly related *Drosophila* species. The *D.virilis* protein has only two TG units tandemly repeated, whereas the corresponding region of the *D.pseudoobscura* protein contains seven TG or TV pairs, interrupted once by the sequence TIIATSG (compare underlined regions in Figure 3). In addition, the short n2 sequence in *D.virilis* shares another feature with that of *D.pseudoobscura*: SGGENSASGS (Figure 3, 678–687) is very similar to SGAENSASGS (830–839). Indeed, the long repeat seen in the *D.pseudoobscura* sequence in n2 appears to be a degenerate version of the first half of this motif.

The remaining non-conserved regions all manifest similarities when compared interspecifically. The n3 regions, while being of quite different lengths, are similar to each other in that they are all predominantly hydrophobic. Yet, only the *D.melanogaster* sequence contains short poly(Gly)

and poly(Ala) stretches like those seen in some other *D. melanogaster* genes (e.g. *Ubx* and *Deformed*) (Laughon *et al.*, 1985; Wilde and Akam, 1987). In *Ubx*, poly(Gly) is present in both the *D.melanogaster* and *D.pseudoobscura* proteins but is of different lengths in the two species. The n4 regions are of similar lengths, and each is quite hydrophilic and lies between two conserved, strongly hydrophobic regions (c4 and c5). Finally, the n5 region of *D.virilis* is longer than those of *D.melanogaster* and *D.pseudoobscura* but is flanked by deletions, at the end of c5 and the beginning of c6, when compared to the *D. melanogaster* and *D.pseudoobscura* proteins. The net result is that, in this region, all three proteins have predominantly hydrophilic stretches of similar lengths at the same approximate positions.

We considered that certain of the non-conserved regions of the *D.pseudoobscura* and *D.virilis* *per* genes might contain known functional motifs that were not detected in protein homology searches performed with the *D.melanogaster* sequence (Jackson *et al.*, 1986; Reddy *et al.*, 1986). However, using the FASTA program of Pearson and Lipman (1988), we have recently searched the National Biomedical Research Foundation (NBRF) protein sequence database (release 16) with the entire sequences of the *D.pseudoobscura* and *D.virilis* *per* proteins. No significant similarity scores were obtained.

Non-coding regions

Only short stretches of sequence were obtained for the 5' and 3' non-coding portions of the *D.pseudoobscura* and *D.virilis* genes (data not shown). The sequence just upstream of the initiating methionine is well conserved in both genes and matches the reported consensus sequence for translational start sites in *Drosophila* genes (Cavener, 1987). In the *D.pseudoobscura* gene, there is an excellent putative 3' splice junction corresponding exactly to the position of the 3' splice junction of the first intron of the *D.melanogaster* gene. An exon of a size appropriate for exon 2 is also detectable in RNase protection experiments with *D. pseudoobscura* RNA (Petersen *et al.*, 1988). Comparable data for the *D.virilis* gene do not exist. Without cDNA clones and more sequence data, we cannot address directly the question of the existence of a first, non-coding exon and a first intron in either the *D.pseudoobscura* or the *D.virilis* gene. As is evident in Figure 1, the introns between the coding exons in the *D.melanogaster* gene are all approximately the same size (57–70 nt), whereas there is much more variability in intron sizes in the *per* genes from the other two species.

Nucleic acid homology searches performed with the intron sequences do not reveal any significant interspecific homology. Presumably, the many insertions, deletions and substitutions make sequence alignments extremely difficult in these regions (cf. Schaeffer and Aquadro, 1987). This result differs from those obtained in comparisons of the *en* gene of *D.melanogaster* and *D.virilis* and the *Ubx* gene of *D.melanogaster* and *D.pseudoobscura*, in which there is some clear homology within introns, probably reflecting a role for these conserved elements in transcription (e.g. Kassis *et al.*, 1986). We suspect that our failure to find interspecific homology within the introns is due to the absence of sequence data for the (putative) first introns of the *D.pseudoobscura* and *D.virilis* *per* genes, given that the large, first intron of

Fig. 2. Comparison of the nucleotide sequences of the coding portions of the *per* genes of *D. virilis*, *D. melanogaster* and *D. pseudoobscura*. The complete nucleotide sequence and its conceptual translation are shown only for the *D. pseudoobscura* (p) *per* gene. The nucleotide sequences of the *per* genes of *D. virilis* (v, this study) and *D. melanogaster* (m, Citri *et al.*, 1987) are shown only where they differ from the *D. pseudoobscura* sequence. Substitutions resulting in amino acid replacements are not highlighted (see Figure 3 for protein sequence comparisons). The conserved regions are shown in triplet codons, whereas the non-conserved regions, n1–n5, are compressed and shown in brackets. The positions of splice junctions are indicated by arrowheads. Deletions relative to the other sequences are signified by a dashed line (the designation of deletions is for comparison purposes only and does not imply the direction or mechanism of change). Numbers are shown for the protein sequence of the *D. pseudoobscura* gene, to facilitate comparison with the protein sequences in Figure 3. Where the *D. pseudoobscura* sequence has a deletion, the translation of the *D. melanogaster* sequence is shown in parentheses. (The sequences of the introns are not presented here but are available upon request.)

the *D. melanogaster per* gene contains information which affects *per* expression, probably at the transcriptional level (Q.Yu *et al.*, unpublished).

Evolutionary mechanisms

When the sequences of the putative *per* genes of three distantly related *Drosophila* species are compared, it is

Fig. 3. Alignment of the protein sequences of the *per* genes of *D. virilis*, *D. melanogaster* and *D. pseudoobscura*. The entire amino acid sequence of the *D. melanogaster per* protein is shown (m); the sequences of the *per* proteins of *D. virilis* (v) and *D. pseudoobscura* (p) are shown only where they differ from the *D. melanogaster* sequence. A dot represents conservation of the amino acid in a given position, whereas a dash represents a deletion. The conserved (c1–c6) and non-conserved (n1–n5) regions are indicated, and the latter are enclosed in brackets. The TG portions of the n2 regions are underlined. (Note that the TG pairs are shown as non-conserved sequences in this figure; however, the detailed nucleotide sequence in Figure 2 shows the extent to which these TG stretches are related at the nucleic acid level.) The numbers on the right refer to the last amino acid in each line. The small brackets within the c2 region encompass the sequences homologous to repeated regions in the *single-minded* gene (cf. Crews *et al.*, 1988).

apparent that 60–70% of the gene has accumulated changes in a manner similar to that observed for a number of other genes. These conserved regions have undergone point mutations resulting in numerous synonymous and replacement substitutions, as well as small deletions and insertions. Overall, these regions of the three proteins have clearly maintained similarity at the amino acid level despite a great many mutational events. In the remaining regions, however, changes have accumulated at more rapid rates, such that the relatedness of these regions is barely discernible; therefore we have designated them ‘non-conserved’. Such regions in the N-terminal half of the protein stand out clearly against a background of some of the most highly conserved portions of the protein. In contrast, both the conserved and non-conserved regions in the C-terminal half of the protein show more variability. Indeed, we suspect that the absence of a clear distinction between conserved and non-conserved regions in the C-terminal portion is primarily responsible for the fact that the extent of amino acid divergence calculated for the conserved regions of *per* gene pairs is

slightly higher than that observed between other gene pairs.

Nucleic acid homology searches, combined with a look at the amino acid sequences and compositions of the non-conserved regions, yield hints concerning some of the mechanisms by which these regions probably evolved. Small stretches of similarity are seen between the three n1 regions and these regions are internally repetitive at the nucleotide level, suggesting a common ancestor that expanded and diverged through a combination of slippage-like mechanisms (Tautz *et al.*, 1986) and point mutation events. The short semi-conserved region near the end of c2 appears to have undergone a deletion in the *D. melanogaster* gene, followed by a duplication of a stretch of nucleotides just downstream, since there is detectable repetitiveness at the nucleic acid level.

The n2 region is especially intriguing, in view of the great variability seen among the three proteins in this region, combined with an obvious regularity of structure in at least two of them. The possible relationship of these features to the function(s) of *per* will be discussed below; however, they

also suggest several mechanisms by which these regions may have diverged. The diminutive *D. virilis* n2 sequence appears to preserve the roots from which the others evolved, perhaps by slippage concurrent with the accumulation of point mutations. This notion is supported by the fact that, within the *D. pseudoobscura* repeat, the degenerate five amino acid repeat units resemble their neighbors more than distant units, suggesting that repeated duplication has occurred more frequently than nucleotide substitutions.

Both the *D. melanogaster* TG repeat region and the *D. pseudoobscura* five amino acid repeat region have experienced extensive intraspecific nucleic acid divergence as evidenced by the number of nucleotide differences among the repeat units. The apparent maintenance of the repeat unit (i.e. the absence of mutational events that destroy the repeat pattern) suggests that the repeats may be under positive selection (cf. Hill and Hastie, 1987). Alternatively, the event that modulates the number of repeat units (e.g. slippage) is much more frequent than events which destroy the repeat. It should be noted that the *D. melanogaster* TG repeat is similar in sequence to other repeated sequences. Probing of Southern blots of genomic DNA from various *Drosophila* and vertebrate species with a TG probe reveals patterns characteristic of repetitive DNA (Shin *et al.*, 1985; H.V. Colot, unpublished observations). Similarly, when a mouse TG-repeat-containing probe was hybridized to DNA from human individuals, the results revealed 'a family of hypervariable minisatellite DNA' (Georges *et al.*, 1987). Although other possibilities exist, we favor the notion that these sequences might be particularly favorable substrates for the slippage-like mechanisms (Tautz *et al.*, 1986) which give rise to tandem direct repeats, both within the *per* gene and elsewhere in genomic DNA.

In other non-conserved regions of the *per* gene there are other, less extensive tandem repeats of amino acids. For example, the n1 region bears a weak relationship to *opa* sequences (Wharton *et al.*, 1985), as the trinucleotide CAX occurs fairly frequently (and sometimes in small clusters) in all reading frames. The *D. melanogaster* n3 region contains poly(Gly) and poly(Ala) stretches. Yet other non-conserved regions contain no clearly discernible repeated sequences. Perhaps other genetic events have given rise to the inter-specific differences in these regions.

per function, conservation and evolution

There is good evidence that regions of a protein important for function are conserved evolutionarily. Indeed, for several enzyme, heat shock and structural protein loci that have been compared among the *Drosophila* species studied here, the amino acid sequences are well conserved (e.g. Blackman and Meselson, 1986; Henikoff and Eghtedarzadeh, 1987; Schaeffer and Aquadro, 1987; Aguade, 1988; Riley, 1989). In this light the relative lack of conservation in the *per* coding region, especially in the 3' half of the gene, is remarkable. This portion includes the *D. melanogaster* TG repeat, whose similarity to a Ser-Gly (SG) repeat of similar length found in some mammalian proteoglycan genes (in which many of the serines serve as sites for the O-linked attachment of glycosaminoglycan (GAG) side chains; Bourdon *et al.*, 1985) suggested that a *per*-encoded proteoglycan might contribute to circadian clock function (Jackson *et al.*, 1986; Reddy *et al.*, 1986; Bargiello *et al.*, 1987). However, single SG dipeptides can serve as sites for GAG chain attachment

(Bourdon *et al.*, 1987). Because there are several TG and SG pairs (but not the TG repeat) in the *per* genes of *D. pseudoobscura* and *D. virilis*, we cannot eliminate the possibility that the *per* protein is a proteoglycan in all three species. In addition, in a strain in which the TG repeat was deleted, two TG pairs were retained and the basic rhythm-generating functions of *per* were intact (Yu *et al.*, 1987a). Further speculations on the relationship of protein modification to *per* function(s) await biochemical characterization of the *per* protein(s).

The 3' portion of the gene is variable at the transcript level even within *D. melanogaster*: the previously described type B and C cDNAs differ from the predominant type A within or near non-conserved blocks in this region. Both type B and C cDNAs are functional (as is type A) in a transformant rescue assay (Citri *et al.*, 1987; J. Ruttila, personal communication). The alternative splicing observed in type B would result in a protein missing most of n3 and the downstream ~40 amino acids, i.e. a protein somewhat similar to the *D. pseudoobscura* and *D. virilis* proteins and, by analogy, perhaps able to perform a subset of, or a slight variation on, the normal function(s) of *per*. On the other hand, it is possible to find acceptable alternative splice consensus sequences [also involving a 3' CG instead of AG (Citri *et al.*, 1987)] in approximately the same positions and in the correct frames in the *D. pseudoobscura* and *D. virilis* genes. The lack of *D. pseudoobscura* and *D. virilis* cDNA clones leaves this issue unresolved.

In the type C transcript, the last three introns are not removed, and the predicted protein would be encoded by an alternative reading frame through intron 5 and into exon 6. There is no apparent conservation, in *D. pseudoobscura* and *D. virilis*, of the predicted primary amino acid sequence at the end of the type C transcript. Furthermore, in neither species is there any open reading frame at all that extends substantially into the corresponding intron. Since neither its specific nor its general features are conserved among the three species studied here, the type C transcript may not play an important part in the conserved functions of *per*, yet it could still encode a protein with functions in *D. melanogaster* that are provided by different minor *per* products (or other proteins) in other species.

The presence of repeated amino acids, and their poor conservation between species, are features shared by the *per* gene with at least two *Drosophila* regulatory loci, *en* and *Ubx* (Kassiss *et al.*, 1986; Wilde and Akam, 1987). In a perhaps related vein, it has recently come to light that some of the functional domains of transcription factors may not be highly conserved, e.g. the 'activating domain' of the yeast positive transcription factor GAL4 can be replaced by a number of other amino acid sequences, related only by charge (Giniger and Ptashne, 1987; Ma and Ptashne, 1987). A related observation is that much of the GAL4 protein is poorly conserved between two yeast species (*Saccharomyces cerevisiae* and *Kluyveromyces lactis*; Salmeron and Johnston, 1986). Finally, we note that both the *per* protein and a *per*- β -gal fusion protein (containing 600 amino acids of *per* sequence at its N terminus) have been localized, at least in part, to nuclei in a number of tissues (Liu *et al.*, 1988; Siwicki *et al.*, 1988). Taken together, these observations suggest that many regulatory proteins in a number of species may contain poorly conserved domains and that *per* may belong to this class of gene products.

The fact that *D.melanogaster* flies, when transformed with the coding region of *D.pseudoobscura per*, manifest behavioral features that resemble those of *D.pseudoobscura* suggests that the *per* coding region contains species-specific instructions (Petersen *et al.*, 1988). In addition, a change in the *D.melanogaster* TG region can have a profound influence on courtship song rhythms (Yu *et al.*, 1987a). Also, recent results (Q.Yu, C.Kyriacou, J.Rutila, D.Wheeler, M.Rosbash and J.C.Hall, unpublished) have shown that the TG repeat region contains information that determines species-specific features of the song rhythm. Based on these observations, we suggest that the variable subregions of the *per* gene contribute significantly to *per* function. Several of these regions are highly variable in primary amino acid sequence between species but appear, at a minimum, to have been under selection for some general structural properties, e.g. hydrophobicity. The *D.melanogaster* TG region is a very imperfect DNA repeat but a perfect amino acid repeat, i.e. no nucleotide substitutions have created amino acid substitutions. The *D.pseudoobscura* five amino acid repeat region, although imperfect, always has a hydrophilic amino acid at every fifth residue. These observations suggest that the repeated portions of the n3 regions, although probably unnecessary for basic *per* function (as they are absent from the *D.virilis* protein), may have some functional significance within a species.

Based on these considerations, we speculate that there are important evolutionary principles that underlie these rather disparate observations. The repeat regions, and more generally the nucleotide sequences that comprise the variable regions of the *per* protein, are hot spots for genetic events that lead to DNA (and therefore protein) changes. Also, these regions of the protein are usually under only modest constraint. Taken together, these two postulates imply that the variable regions give rise to diversity within a population by generating structurally and functionally diverse *per* genes. This should, in turn, give rise to a behaviorally diverse population. As *per* influences the male courtship song rhythm (e.g. Kyriacou and Hall, 1986; Yu *et al.*, 1987b) and is likely, therefore, to influence courtship, mating and reproductive success, we suggest that mutational events within the *per* gene may contribute directly to speciation.

Materials and methods

Southern analysis and subcloning of per fragments

Genomic DNA was isolated as described (Bender *et al.*, 1983), digested with *Eco*RI and *Bam*HI, and analyzed on Southern blots by standard techniques (Maniatis *et al.*, 1982), except that the hybridization mix contained 10 × Denhardt's, 30% formamide and 250 µg/ml salmon sperm DNA and was incubated at 37°C ('low stringency'). Additional Southern blots were done with *Cfo*I-digested genomic DNA and denaturing acrylamide gels as described by Yu *et al.* (1987a). All probes (see text) were labeled with random primers and Klenow fragment (Feinberg and Vogelstein, 1983). For the cloning of small *per*-homologous fragments, *Hin*PI-digested genomic DNA was fractionated on a non-denaturing acrylamide gel alongside a 123 bp marker (BRL). DNA fragments in the proper size fraction (as previously determined from the blots of *Cfo*I-digested DNA) were eluted and purified by standard techniques. The fragments were then shotgun-cloned into *Acc*I-digested pEMBL18+ (Dente *et al.*, 1983); the resulting colonies were screened with the same probes and under the same low stringency conditions as the Southern blots (above), and single-stranded DNA was prepared from positive clones by standard techniques and sequenced by the dideoxy chain termination method (Sanger *et al.*, 1977).

Isolation and analysis of genomic per clones

The *D.pseudoobscura* library was a generous gift of Chuck Langley and Celeste Brown, and was made by cloning *Mbo*I partially digested DNA into the *Bam*HI site of EMBL4. The *D.virilis* library, kindly provided by Ron Blackman, was similarly constructed with *Mbo*I partially digested DNA and EMBL3. The subclones of the *Hin*PI fragments generated above were used to screen their corresponding genomic phage libraries by standard techniques. Approximately one positive plaque was obtained per genome screened. DNA was prepared from five positive phage from each library and analyzed by digestion with *Eco*RI. Some of the same digests were shotgun-cloned into pTZ18U (Mead *et al.*, 1986) and the colonies screened with the appropriate *Hin*PI fragment subclones. Single-stranded DNA from positive clones was used to generate nested deletions by the method of Dale *et al.* (1985), which were then sequenced with the Sequenase kit (US Biochemical Corp.) and buffer gradient polyacrylamide gels (Biggin *et al.*, 1983). Sequence data were analyzed by the programs of Pustell and Kafatos (1986) and by the SYNSUB program of Lewontin (1989). A search of the NBRF protein sequence database was performed with the FASTA program of Pearson and Lipman (1988).

Acknowledgements

We are grateful to Chuck Langley, Celeste Brown and Ron Blackman for their generous gifts of phage libraries; to Steve Schaeffer for help with data analysis; to Peg Riley, Marty Kreitman, George Miklos, Alec Jeffreys and Charalambos Kyriacou for comments on the manuscript; and to Tobie Tishman for typing it. The work was supported by a grant from the NIH (GM-33205) to M.R. and J.C.H.

References

- Agua, M. (1988) *Mol. Biol. Evol.*, **5**, 433–441.
- Bargiello, T.A., Saez, L., Baylies, M.K., Gasic, G., Young, M.W. and Spray, D.C. (1987) *Nature*, **328**, 686–691.
- Baylies, M.K., Bargiello, T.A., Jackson, F.R. and Young, M.W. (1987) *Nature*, **326**, 390–392.
- Bender, W., Spierer, P. and Hogness, D.S. (1983) *J. Mol. Biol.*, **168**, 17–33.
- Biggin, M.D., Gibson, T.J. and Hong, G.F. (1983) *Proc. Natl. Acad. Sci. USA*, **80**, 3963–3965.
- Blackman, R.K. and Meselson, M. (1986) *J. Mol. Biol.*, **188**, 499–515.
- Bourdon, M.A., Oldbert, A., Pierschbacher, M. and Ruoslahti, E. (1985) *Proc. Natl. Acad. Sci. USA*, **82**, 1321–1325.
- Bourdon, M.A., Krusius, T., Campbell, S., Schwartz, N.B. and Ruoslahti, E. (1987) *Proc. Natl. Acad. Sci. USA*, **84**, 3194–3198.
- Cavener, D.R. (1987) *Nucleic Acids Res.*, **15**, 1353–1361.
- Crews, S.T., Thomas, J.B. and Goodman, C.S. (1988) *Cell*, **52**, 143–151.
- Citri, Y., Colot, H.V., Jacquier, A.C., Yu, Q., Hall, J.C., Baltimore, D. and Rosbash, M. (1987) *Nature*, **326**, 42–47.
- Dale, R.M.K., McClure, B.A. and Houchins, J.P. (1985) *Plasmid*, **13**, 31–40.
- Dente, L., Cesarini, G. and Cortese, R. (1983) *Nucleic Acids Res.*, **11**, 1645–1654.
- Feinberg, A.P. and Vogelstein, B. (1983) *Anal. Biochem.*, **132**, 6–13.
- Flavell, A.J., Dyson, J. and Ish-Horowicz, D. (1987) *Nucleic Acids Res.*, **15**, 4035–4048.
- Georges, M., Cochaux, P., Lequarre, A.-S., Young, M.W. and Vassart, G. (1987) *Nucleic Acids Res.*, **15**, 7193.
- Giniger, E. and Ptashne, M. (1987) *Nature*, **330**, 670–672.
- Hall, J.C. and Roshash, M. (1988) *Annu. Rev. Neurosci.*, **11**, 373–393.
- Henikoff, S. and Eghtedarzadeh, M.K. (1987) *Genetics*, **117**, 711–725.
- Hill, R.E. and Hastie, N.D. (1987) *Nature*, **326**, 96–99.
- James, A.A., Ewer, J., Reddy, P., Hall, J.C. and Rosbash, M. (1986) *EMBO J.*, **5**, 2313–2320.
- Jackson, F.R., Bargiello, T.A., Yun, S.-H. and Young, M.W. (1986) *Nature*, **320**, 185–188.
- Kassis, J.A., Poole, S.J., Wright, D.K. and O'Farrell, P.H. (1986) *EMBO J.*, **5**, 3583–3589.
- Kyriacou, C.P. and Hall, J.C. (1986) *Science*, **232**, 494–497.
- Laughon, A., Carroll, S.B., Storfor, F.A., Riley, P.D. and Scott, M.P. (1985) *Cold Spring Harbor Symp. Quant. Biol.*, **50**, 253–262.
- Lewontin, R.C. (1989) *Mol. Biol. Evol.*, in press.
- Liu, X., Lorenz, L., Yu, Q., Hall, J.C. and Rosbash, M. (1988) *Genes Dev.*, **2**, 228–238.
- Ma, J. and Ptashne, M. (1987) *Cell*, **51**, 113–119.

- Maniatis,T., Fritsch,E.F. and Sambrook,J. (1982) *Molecular Cloning. A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
- Mead,D.A., Szczesna-Skorupa,E. and Kemper,B. (1986) *Prot. Eng.*, **1**, 67–74.
- Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.
- Petersen,G., Hall,J.C. and Rosbash,M. (1988) *EMBO J.*, **7**, 3939–3947.
- Pustell,J. and Kafatos,F.C. (1986) *Nucleic Acids Res.*, **14**, 479–488.
- Reddy,P., Zehring,W.A., Wheeler,D.A., Pirrotta,V., Hadfield,C., Hall,J.C. and Rosbash,M. (1984) *Cell*, **38**, 701–710.
- Reddy,P., Jacquier,A.C., Abovich,N., Petersen,G. and Rosbash,M. (1986) *Cell*, **46**, 53–61.
- Riley,M.A. (1989) *Mol. Biol. Evol.*, in press.
- Salmeron,J.M., Jr and Johnston,S.A. (1986) *Nucleic Acids Res.*, **14**, 7767–7781.
- Sanger,R., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA*, **74**, 5463–5467.
- Schaeffer,S.W. and Aquadro,C.F. (1987) *Genetics*, **117**, 61–73.
- Shin,H.-S., Bargiello,T.A., Clark,B.T., Jackson,F.R. and Young,M.W. (1985) *Nature*, **317**, 445–448.
- Siwicki,K.K., Eastman,C., Petersen,G., Hall,J.C. and Rosbash,M. (1988) *Neuron*, 141–150.
- Tautz,D., Trick,M. and Dover,G.A. (1986) *Nature*, **322**, 652–656.
- Wharton,K.A., Yedvobnick,B., Finnerty,V.G. and Artavanis-Tsakonas,S. (1985) *Cell*, **40**, 55–62.
- Wilde,C.D. and Akam,M. (1987) *EMBO J.*, **6**, 1393–1401.
- Yu,Q., Colot,H.V., Kyriacou,C.P., Hall,J.C. and Rosbash,M. (1987a) *Nature*, **326**, 765–769.
- Yu,Q., Jacquier,A.C., Citri,Y., Hamblen,M., Hall,J.C. and Rosbash,M. (1987b) *Proc. Natl. Acad. Sci. USA*, **84**, 784–788.
- Zehring,W.A., Wheeler,D.A., Reddy,P., Konopka,R.J., Kyriacou,C.P., Rosbash,M. and Hall,J.C. (1984) *Cell*, **39**, 369–376.

Received on July 26, 1988; revised on September 9, 1988

Note added in proof

Consistent with the implication that there is one *per*-like gene in the *D.pseudoobscura* genome (see above), use of the DNA fragment diagrammed in Figure 2 (middle) of Petersen *et al.* (1988) in larval salivary gland *in situ* hybridization experiments showed one site of labelling distally on the left arm of *D.pseudoobscura*'s X chromosome.