Molecular Evolution of a Repetitive Region within the *per* Gene of *Drosophila*

Alexandre A. Peixoto,* Susanna Campesan,† Rodolfo Costa,† and Charalambos P. Kyriacou*

*Department of Genetics, University of Leicester; and †Dipartimento di Biologia, Università di Padova

The clock gene *period* (*per*) controls a number of biological rhythms in *Drosophila*. In *D. melanogaster*, *per* has a repetitive region that encodes a number of alternating threonine-glycine residues. We sequenced and compared this region from several different *Drosophila* species belonging to various groups within the *Drosophila* and *Sophophora* subgenera. This part of *per* shows a great variability in both DNA sequence and length. Furthermore, analysis of the data suggests that changes in the length of this variable region might be associated with amino acid replacements in the more conserved flanking sequences.

Introduction

The *period* (*per*) gene in *Drosophila melanogaster* determines biological rhythmicity in a number of temporal domains, including the circadian (the 24-h locomotor activity cycle; Konopka and Benzer 1971), the ultradian (the 60s love-song rhythm; Kyriacou and Hall 1980, 1989), and the infradian (egg-to-adult developmental cycle; Kyriacou et al. 1990). In *D. melanogaster*, the primary translation product of the *per* gene is a protein of ~1,200 amino acids, whose most prominent feature is a stretch of alternating threonine-glycine (Thr-Gly) residues (Jackson et al. 1986; Citri et al. 1987). The number of Thr-Gly encoding repeats varies both within and between the species of the *melanogaster* subgroup (Yu et al. 1987; Thackeray and Kyriacou 1990; Costa et al. 1991; Peixoto et al. 1992).

The Thr-Gly repeat in *D. melanogaster* may be involved in the thermostability of the circadian phenotype, because removing it from the *per* gene and transforming arrhythmic *per*<sup>01</sup> mutants with the deleted construct gives a temperature-sensitive circadian phenotype (Ewer et al. 1990). This is in marked contrast to the results obtained with transformants carrying a normal *per*<sup>+</sup> gene. Furthermore, short Thr-Gly variants are found in higher frequencies in southern Europe than in northern Europe, where longer Thr-Gly repeat alleles predominate (Costa et al. 1992). Latitudinal clines traditionally suggest that selection may be important in maintaining polymorphisms (e.g., see Anderson and Oakeshott 1984; David et al. 1989). However, in our case it cannot be excluded that selection is acting on sequences closely linked to the Thr-Gly repeat.

Recent comparisons of the Thr-Gly sequences between the species within the *melanogaster* subgroup show a rapid evolution in this region by a combination of

1. Key words: *Drosophila*, period gene, repetitive DNA, biological rhythms.

Address for correspondence and reprints: Charalambos P. Kyriacou, Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, United Kingdom.

© 1993 by The University of Chicago. All rights reserved.
0737-4038/93/1001-0008$02.00
slippage-like events and point mutations (Peixoto et al. 1992). However, the variation in length of the Thr-Gly region among the eight species of the subgroup is surprisingly small, given the potential for length mutations that has been observed within *D. melanogaster* (Costa et al. 1991). This implies that there must be some constraint acting on the length of this region within the subgroup.

The complete *per* coding sequences of *D. pseudoobscura, D. virilis* (Colot et al. 1988), and *D. yakuba* (Thackeray and Kyriacou 1990) have also been reported. A general feature appears to be that the 3' part of the gene is more variable between species than is the 5' part. In *D. pseudoobscura* the Thr-Gly repeat is largely replaced by a five-amino-acid degenerate motif (Colot et al. 1988), which is also polymorphic in length (Costa et al. 1991), whereas *D. virilis* has a very short Thr-Gly region (Colot et al. 1988).

In the present paper, we have sequenced the region corresponding to the “Thr-Gly” repeat in a number of other species from the *Drosophila* genus. The comparison of these sequences shows an area of high variability, both in length and sequence composition. However, the flanking regions are more conserved, and changes in length of the diverged region might possibly be associated with amino acid changes in the flanking sequences.

**Material and Methods**

**Drosophila Strains**

The following *Drosophila* strains (stock numbers are in parentheses) were used in this work: *D. ananassae* (0371.0), *D. immigrans* (1731.0), *D. mediostriata* (2391.0), *D. mojavensis*, *D. robusta* (1111.1), *D. saltans* (0911.0), *D. serrata* (0681.0), and *D. willistoni* (0811.0). With the exception of *D. mojavensis* (obtained from Dr. E. Zouros), all other strains were obtained from the Bowling Green State University, Ohio.

**PCR Amplification and DNA Sequencing**

The sequences were obtained by PCR and direct DNA sequencing according to a method described elsewhere (Costa et al. 1991). In brief, the PCR amplification was carried out for 30 cycles (95°C for 1 min, 65°C or 60°C for 1 min, and 70°C for 1 min) in a Perkin Elmer Cetus thermocycler. Fly DNA used in the PCR reactions was prepared using the method of Gloor and Engels (1990). Single males were ground in 50 μl of buffer (10 mM Tris-HCl pH 8.2, 1 mM EDTA, 25 mM NaCl, 200 μg proteinase K/ml), were left for 30 min at 37°C, and then were heated above 95°C for 2 min. One microliter of genomic DNA was used for each 10 μl of PCR reaction. The DNA sequencing was carried out by using the Sequenase version 2.0 kit from United States Biochemical. The primers used for the PCR or sequencing were the following ones: 5' primers—5'-CCCGTCCACGAGGGCAGGGGGG-3' (5005–5028), 5'-AAC-TATAACGAGAACCTGCT-3' (4874–4893), 5'-GGCAGCAATGTGCATGAG-3' (5057–5076), and 5'-GGCAGCAATATACACATGAG-3' (5057–5076), and 3' primers—5'-CCGCGCGACTGCCGGGTCTCTTCA-3' (5364–5387) and 5'-TTTCTCATCTCGCTTGTTGTT-3' (5336–5355). The primer positions (in parentheses) refer to the *D. melanogaster* sequence published by Citri et al. (1987), and different combinations of primers were used, depending on the species.

**Analysis of the DNA and Protein Sequences**

The sequences obtained were analyzed using the University of Wisconsin Genetics Computer Group (UWGCG) software (version 5.0; Devereux et al. 1984), the program
Results

Figure 1 shows the amino acid sequences of the putative per protein in the conserved flanking areas that surround the variable-length region. The sequences from 11 different Drosophila species belonging to both the Drosophila and Sophophora subgenera are illustrated. The sequence from D. melanogaster is taken from Citri et al. (1987), and the sequences of D. virilis and D. pseudoobscura are from Colot et al. (1988). The alignments are ordered according to the length of the centrally located variable region. The flanking regions are defined simply as those that can be easily aligned and that do not show any evidence for slippage-like events among the species compared. Most of the amino acid changes in the flanking conserved regions involve amino acids with similar characteristics. There are enormous between-species differences in length of the variable region.

The DNA sequences of the flanking conserved regions are shown in figure 2. These data were used to construct an unrooted tree for the Drosophila species by using Kimura's (1980) two-parameter distance and the neighbor-joining method (Saitou and Nei 1987), which is illustrated in figure 3. Bootstrapping shows none of the interior branches to be statistically significant. However, this test is probably too severe for phylogenies based on short sequences (Pelandakis et al. 1991). Nevertheless, the clustering of the species in the Drosophila and Sophophora subgenera is in general agreement with the results of Throckmorton (1975). The only unusual feature is the position of D. ananassae and D. serrata, which are not clustered together with D. melanogaster. These species represent the ananassae and monium subgroups of the melanogaster group (Lemeunier et al. 1986) and were therefore expected to have a common root with D. melanogaster. However, our results are similar to those obtained by Pelandakis et al. (1991) using rRNA sequences.

Figure 4 shows the DNA and protein sequences in the variable-length region for

<table>
<thead>
<tr>
<th>N-terminal conserved region</th>
<th>variable length region</th>
<th>C-terminal conserved region</th>
</tr>
</thead>
<tbody>
<tr>
<td>vir</td>
<td>EGSGGGSGSGNLTTASNVRMSSVTNTSNTGTG</td>
<td>19</td>
</tr>
<tr>
<td>med</td>
<td>F.G.</td>
<td>-22-</td>
</tr>
<tr>
<td>imm</td>
<td>F.G.</td>
<td>-22-</td>
</tr>
<tr>
<td>moj</td>
<td>F.G.</td>
<td>-24-</td>
</tr>
<tr>
<td>sal</td>
<td>F.G.</td>
<td>-25-</td>
</tr>
<tr>
<td>wil</td>
<td>F.G.</td>
<td>-27-</td>
</tr>
<tr>
<td>rob</td>
<td>F.G.</td>
<td>-45-</td>
</tr>
<tr>
<td>mel</td>
<td>F.G.</td>
<td>-66-</td>
</tr>
<tr>
<td>ana</td>
<td>F.G.</td>
<td>-84-</td>
</tr>
<tr>
<td>ser</td>
<td>H.G.</td>
<td>-100-</td>
</tr>
<tr>
<td>pse</td>
<td>A., H.S.G.</td>
<td>-209-</td>
</tr>
</tbody>
</table>

**Fig. 1.—** Alignment of the amino acid sequences in the conserved flanking regions of 11 different Drosophila species: D. virilis (vir), D. mediodiastara (med), D. immigrans (imm), D. mojavensis (moj), D. saltans (sal), D. willistoni (wil), D. robusta (rob), D. melanogaster (mel), D. ananassae (ana), D. serrata (ser), and D. pseudoobscura (pse). The sequence of D. melanogaster is from Citri et al. (1987), and the sequences of D. virilis and D. pseudoobscura are from Colot et al. (1988). The sequences were aligned using the program CLUSTAL V (Higgins and Sharp 1988). The stars and dots in the bottom line represent, respectively, identical and similar amino acids positions.
FIG. 2.—Alignment of the DNA sequences in the conserved flanking regions. The stars represent identical nucleotide positions. Species are represented by the same symbols as in fig. 1.
Repetitive Region of *Drosophila per* Gene

The 11 species. The species differ considerably in both sequence and length (also see fig. 1). Moreover, each species shows segments of perfect or imperfect repeats. These are highlighted by arrows and lines above the sequences. The sequences clearly diverged by a combination of point mutations and slippage-like events. For example, the SGGEN protein motif in *D. virilis* is mutated to SGGGN in *D. mediostriata* and *D. immigrans* to SAGGN in *D. mojavensis*, where it is duplicated and mutated to the motif SAGDN. In *D. pseudoobscura* this five-amino-acid motif is mutated and greatly amplified, giving rise to a large degenerate repeat region. The direction of the mutational events suggested above is arbitrary. Figure 4 shows how the Thr-Gly repeats appear to degenerate downstream of the main repeat in *D. melanogaster*, *D. ananassae*, and *D. serrata*.

We investigated possible constraints on the evolution of this region, in spite of the high divergence. For example, according to secondary-structure predictions, the variable region in *D. melanogaster* and *D. pseudoobscura* represents a region of turns that appears to divide the *per* protein into two domains (Costa et al. 1991). We examined the predicted secondary structure for this region in the other nine species by using the method of Chou and Fasman (1974). As with *D. melanogaster* and *D. pseudoobscura* (Costa et al. 1991), a region of turns was usually predicted (data not shown), in spite of the heterogeneity in the amino acid composition. Figure 5 shows an alignment of the complete amino acid sequence for two pairs of species that have relatively similar lengths of the variable region. *Drosophila mojavensis* and *D. saltans*, which belong to different subgenera, are quite similar in this region. The presence of long Thr-Gly repeats in both *D. melanogaster* and *D. robusta*, which also belong to different subgenera, is interesting. Although it is possible that the presence of a long
FIG. 4.—DNA and protein sequences of the variable-length region among different species of the *Drosophila* genus (also see legend to fig. 1). The arrows and lines above the sequences highlight perfect and imperfect repeated DNA and amino acid motifs.
Fig. 4 (Continued)

moj  EGSGGSGSSGNFTGNSVMSSVTNTSNAAGTGTASGAANDAAAGGSGRNSAPAVT---VTLLLEDVNLK
sal  EGSSGSGSSGNFTGNSVMSSVTNTSNAAGTGTASGAANDAAAGGSGRNSAPAVT---VTLLLEDVNLK

rob  EGSSGSGSSGNFTGNSVMSSVTNTS---NAAGTGTGTGTGTGT GTGTSAGGTSAGGNANSGN---GNPFPAFAITTLLEDVNLK
me1  EGSSGSGSSGNFTGNSVMSSVTNTSNAAGTGTGTGTGTGT GTGTSAGGTSAGGNANSGN---GNPFPAFAITTLLEDVNLK

Fig. 5.—Alignment of the complete amino acid sequence between two pairs of species that have relatively similar lengths of the variable region. As in fig. 1, the stars indicate identical amino acids, while the dots indicate similar amino acid positions.
Thr-Gly repeat in both species may reflect an ancestral state, it seems more probable, when one considers the distant phylogenetic relationships of *D. robusta* from *D. melanogaster* (Throckmorton 1975; Rousset et al. 1991), that this is a case of convergent evolution at the molecular level.

The repetitive region of *per* has features similar to minisatellite sequences, which also show large differences in length when different species are compared (Gray and Jeffreys 1991). Because of the rapid evolutionary rate of these sequences, one does not expect to find a correlation between the time of separation of a pair of species and any length difference between their sequences. In the case of the *per* gene, it is unlikely that the repetitive region is totally free of selective constraints. Consequently we might predict that large between-species differences in the length of these sequences should not be observed. This prediction is confirmed by examining the differences in length of the Thr-Gly region within the *melanogaster* subgroup of species, differences that range from 47 amino acids in *D. teissieri* to 76 amino acids in *D. mauritiana* (Peixoto et al. 1992). This relatively small range is surprising, given that a single slippage-like event within *D. melanogaster* can alter the length of this region by 18 amino acids (Costa et al. 1991).

In marked contrast to these results, much larger differences in length are observed, in the corresponding region, between the more distantly related species (see figs. 1 and 4). This difference in length can be as large as one order of magnitude (19 amino acids in *D. virilis* vs. 209 amino acids in *D. pseudoobscura*). Thus the question arises as to whether these interspecific length differences correlate with the time of divergence between these species.

We therefore examined whether the divergence in length of the variable region was correlated with the divergence in the DNA sequences of the conserved flanking regions. In the upper graph of figure 6 the percentage of third-base nucleotide differences between pairs of species in the conserved flanking regions (most of which are synonymous and therefore likely to be selectively neutral) is plotted against the corresponding pairwise difference in length of the protein in the variable region. The correlation is not significant (Spearman's rho = 0.1019; N = 55; not significant; all tests are one-tailed, as the correlation is not expected to be negative). The correlation is even smaller if the third-base data are corrected for multiple hits (rho = 0.055; Kimura 1980) or if the proportion of synonymous differences is used instead (rho = 0.0338; Nei and Gojobori 1986). However, when the pairwise differences in length of the protein in the variable region are plotted against the percentage of amino acid differences in the conserved flanking regions, the correlation is highly significant (fig. 6, lower graph; rho = 0.5078; N = 55; P < 0.001). The same results were obtained using the Mantel test (Mantel 1967), a matrix correlation test.

If the synonymous mutation rate represents the ticking of the molecular clock, then the lack of correlation between the third-base position and the pairwise length difference indicates that the latter is not evolving in a clocklike manner. Therefore, the positive correlation observed between the pairwise length difference and the protein divergence is probably not simply due to the length of time since the divergence between the species. Alternatively, perhaps the third-base positions in the flanking region are close to saturation by mutation, rendering them uninformative. To address this question we correlated our third-base data with the pairwise percentage divergence data based on published rRNA sequences available for 8 of the 11 species (Rousset et al. 1991). The correlation was significant (rho = 0.4412; N = 28; P < 0.02), indicating that the third-base position in our data is still informative. Furthermore, when
the same rRNA data were compared with our protein data, the correlation was not higher than that obtained with the third-base data (rho = 0.3306), demonstrating that the protein is not simply more informative than the third-base position.

Care has to be taken when interpreting correlations derived from comparison between pairs of species, as they cannot be considered independent points, because of the phylogenetic relationships between them (Felsenstein 1985b). In our case, however, the lack of significant correlation observed for the third-base-position data may
act as an internal control. The bias that could favor a positive correlation between pairwise length differences and protein divergence in the conserved region should also be seen in the correlation between the former and the DNA divergence based on the third base. Nevertheless, we repeated the correlations after first performing a phylogenetic correction of our data, as do Coyne and Orr (1989); but also see the work of Felsenstein (1985b). We used the phylogeny based on the rRNA data mentioned above (Rousset et al. 1991). For the three species that were not presented by Rousset et al., the positions in the tree of the closest known relative were used. The phylogenetic correction reduced our data set to only 10 points. However, the correlation between the length differences in the variable region and the protein divergence in the conserved region was still significant \( \rho = 0.6809; N = 10; P < 0.05 \), whereas the correlation with the third-base DNA divergence was not significant \( \rho = 0.3222 \), confirming our previous results.

Discussion

We have shown that the region of \textit{per} corresponding to the Thr-Gly repeat in \textit{Drosophila melanogaster} has diverged considerably, in both length and amino acid composition, among these 11 species. Furthermore, the correlations suggest that changes in length of the diverged region could conceivably be associated with amino acid changes in the flanking sequences. However, it is premature to say that our results present evidence for intragenic molecular coevolution, as there may be additional complications associated with the correlations. For instance, although unlikely, the lack of correlation with the third base could be the result of between-species differences in the codon usage patterns (Starmer and Sullivan 1989). Also, perhaps, the phylogenetic correction that we borrowed from Coyne and Orr (1989) might not be entirely appropriate for our data set (J. Felsenstein, personal communication). However, given that both uncorrected and corrected phylogenetic analysis gave similar results regarding the lack of any significant correlation of the length differences with the third-base position, as opposed to the significant length-versus-protein-divergence relationship, we feel more confident that we may be seeing a real effect. Supporting our views is the comparison of this region of \textit{per} in two very distantly related species of dipterans from the genera \textit{Lucilia} and \textit{Sarcophaga}. These two species have very short and almost identical variable regions but reveal only one amino acid change in the conserved flanking sequences (A. Branscombe and R. White, personal communication). This represents the same number of amino acid replacements in the conserved region as there are between the \textit{erecta-orena} lineage and the other species of the \textit{melanogaster} subgroup, where relatively small length differences occur. Therefore, we suspect that, given a certain time of divergence, species with more-similar-sized variable regions will tend to have more-similar flanking sequences.

It has been proposed that genes affected by genetic turnover mechanisms such as slippage, mechanisms that usually occur with higher frequencies than do point mutations, could drive other changes in interacting genes (Dover and Tautz 1986). The high mutational pressure associated with slippage-like events in the “Thr-Gly” domain might possibly drive further changes, such as compensatory mutations, in other parts of the protein and thus accelerate the rate of evolution of \textit{per}. Kimura (1991) has shown that a combination of very tightly linked compensatory neutral mutations can easily become fixed in a population by random drift even when the individual mutations are deleterious. This process would be enhanced by a region associated with high mutation rates, such as the variable region of \textit{per}. It is also possible
that adaptive changes in the flanking sequences could change the constraints imposed on the length of the variable region. Alternatively, adaptive changes in the length of the variable region could be followed by selectively driven compensatory mutations in the conserved flanking regions.

The *per* region appears to play an important functional role in modulating different types of biological cycles (Yu et al. 1987; Ewer et al. 1990; Wheeler et al. 1991). By mixing and matching flanking and diverged regions between the species and by following this with the relevant behavioral analysis of transformants, it may be possible to detect whether certain combinations of amino acids in the two regions have co-evolved. A convincing demonstration of intragenic molecular coevolution will ultimately have to rely on such experimental manipulations. Our preliminary analysis of the sequence data represents only a tentative suggestion that such a coevolutionary process may be occurring in this part of the *per* gene.

Finally, in a rapidly evolving region of the *per* gene, such as the “Thr-Gly” domain, which constraints might be acting on the per protein? This is not yet very clear. One possibility comes from the secondary-structure predictions, which suggest that, in spite of the interspecific variability in amino acid composition, a region of turns is generally present. Moreover, the variable-length region in all the species is mostly composed of small amino acids (Taylor 1986). Another possible indication that constraints are affecting the evolution of this region comes from the high level of similarity that we observed between pairs of species belonging to different subgenera (*D. mojavensis–D. saltans* and *D. robusta–D. melanogaster*; fig. 5). These findings suggest that the variable region can only evolve toward a relatively small number of different amino acid motifs, leading to evolutionary convergence. It is interesting that the only region of similarity between the *D. melanogaster per* gene and the *Neurospora* clock gene frequency (frq) is in the area including and surrounding the Thr-Gly repeats (McClung et al. 1989). The implication is that this region may play an important general role in these two clock proteins.

Sequence Availability

Sequences have been deposited in GenBank under the following accession numbers: L06335, *Drosophila ananassae* (375 bp); L06336, *D. saltans* (198 bp); L06337, *D. immigrans* (189 bp); L06338, *D. mediostriata* (189 bp); L06339, *D. mojavensis* (195 bp); L06340, *D. robusta* (258 bp); L06341, *D. serrata* (423 bp); and L06342, *D. willistoni* (204 bp).

Acknowledgments

This study was supported by an SERC grant to C.P.K., by a CEC grant to C.P.K. and R.C., by a MURST-British Council award for Anglo-Italian cooperation, by a Brazilian CNPq scholarship to A.A.P., and by an ERASMUS studentship to S.C. We thank John Brookfield, Gabby Dover, Joe Felsenstein, Doug Futuyma, Wen-Hsiung Li, Masatoshi Nei, Paul Sharp, and Tom Whittam for their comments on earlier versions of the manuscript. We also thank Koichiro Tamura for redrawing figure 3 for us.

LITERATURE CITED


MASATOSHI NEI, reviewing editor

Received June 12, 1992; revision received August 17, 1992

Accepted August 18, 1992