Patterns of Nucleotide Polymorphism and Divergence in the Odorant-Binding Protein Genes *OS-E* and *OS-F*: Analysis in the Melanogaster Species Subgroup of Drosophila

Alejandro Sánchez-Gracia, Montserrat Aguadé and Julio Rozas[1]

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain

Manuscript received April 11, 2003
Accepted for publication July 25, 2003

ABSTRACT

The *Olfactory Specific-E* and *-F* genes (*OS-E* and *OS-F*) belong to the odorant-binding protein gene family, which includes the general odorant-binding proteins and the pheromone-binding proteins. In *Drosophila melanogaster*, these genes are arranged in tandem in a genomic region near the centromere of chromosome arm 3R. We examined the pattern of DNA sequence variation in an ~7-kb genomic region encompassing the two *OS* genes in four species of the melanogaster subgroup of Drosophila and in a population sample of *D. melanogaster*. We found that both the *OS-E* and the *OS-F* gene are present in all surveyed species. Nucleotide divergence estimates would support that the two genes are functional, although they diverge in their functional constraint. The pattern of nucleotide variation in *D. melanogaster* also differed between genes. Variation in the *OS-E* gene region exhibited an unusual and distinctive pattern: (i) a relatively high number of fixed amino acid replacements in the encoded protein and (ii) a peak of nucleotide polymorphism around the *OS-E* gene. These results are unlikely under the neutral model and suggest the action of natural selection in the evolution of the two odorant-binding protein genes.

THE olfactory system of terrestrial animals has an extreme sensitivity and specificity. It can detect and discriminate a large number of olfactory signals, the odorants. Olfactory perception is accomplished by specialized bipolar sensory neurons that extend their dendrites into an aqueous medium: the olfactory mucus in vertebrates and the sensillar fluid in insects (Stern and Marx 1999). Hence, the airborne molecules must traverse the aqueous space that separates neuronal cells from the external air and stimulate the odorant receptors (Steinbrecht 1969, 1996). These receptors are located on the dendritic membrane of the sensory neurons (Buck and Axel 1991; Vosshall *et al.* 1999).

The odorant-binding proteins (OBPs) are abundant low-molecular-weight proteins that bind and solubilize hydrophobic odorants (or pheromones) in the vertebrate olfactory mucus and in the insect sensillar lymph. These small globular proteins are synthesized and secreted by some accessory cells surrounding the sensory neurons. In insects, the OBP family includes the general odorant-binding proteins (GOBPs) and the pheromone-binding proteins (PBPs), which are not homologous to vertebrate odorant-binding proteins (Vogt and Riddiford 1981; Pelosi and Maida 1995).

Despite the low sequence similarity among different insect OBPs, most of these proteins exhibit a similar distribution of conserved hydrophobic residues with a nearly identical predicted secondary structure. Most proteins of this family contain six highly conserved cysteines located in similar positions of the protein (Pikielny *et al.* 1994). In Lepidoptera, these cysteines are involved in disulfide bridges in both PBPs and GOBPs (Scaloni *et al.* 1999). The similar distribution of cysteine residues in both groups of OBPs suggests that the disulfide-bridge pairing might be a general feature of this family of molecules in insects.

Although the specific function of OBPs in olfaction is still unknown, they seem to play an important role in olfactory coding. It has been shown that several OBPs have different odorant specificities and are present in distinct subsets of antennal sensilla (Pelosi and Maida 1995). Additionally, genes encoding olfactory receptors with different binding specificities are also expressed in specific areas of the olfactory organ (Vosshall *et al.* 2000). These observations suggest that these proteins might participate in odor detection by restricting the spectrum of odorants accessible to the underlying receptors. In addition to the established functions of OBPs as carrier molecules and in concentrating hydrophobic odorants in the aqueous medium, it has also been proposed that these proteins could participate in the deactivation of the odorant stimulus (Pelosi and Maida 1995).

In *Drosophila melanogaster*, 51 putative members of the OBP family have been identified (Hekmat-Scafe *et al.* 2002; Vogt *et al.* 2002). Two of theses proteins, OS-E

Sequence data from this article have been deposited in the EMBL/GenBank Data Libraries under accession nos. AJ574644, AJ574762–AJ574774 (*D. melanogaster*), AJ563750 (*D. mauritiana*), AJ567753 (*D. simulans*), and AJ574775–AJ574776 (*D. erecta*).

[1]*Corresponding author:* Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Diagonal 645 08028, Barcelona, Spain. E-mail: jrozas@ub.edu

(olfactory-specific E) and OS-F (olfactory-specific F), colocalize at the same restricted area of the ventrolateral region of the antenna. The *OS-E* and *OS-F* genes (named *Obp83a* and *Obp83b* in Hekmat-Scafe *et al.* 2002), which have a similar intron-exon organization, are arranged in tandem in cytological band 83CD of the third chromosome. The two encoded proteins are highly conserved (72% amino acid identity in the mature protein; Pelosi and Maida 1995) except for a region in the C-terminal domain of the proteins, which has been named the heterogeneous region (*hr*; Hekmat-Scafe *et al.* 2000). The close physical proximity and the high degree of sequence similarity of the two coding regions seem to reflect a recent gene duplication event (McKenna *et al.* 1994; see, however, Hekmat-Scafe *et al.* 2000). Two hypotheses have been proposed to explain the low amino acid conservation of the *hr* region: (i) the *hr* region might form a putative binding site for the odorant molecule and (ii) the *hr* region might be a putative contact site for the olfactory receptor (Hekmat-Scafe *et al.* 2000).

Since olfaction is essential for survival and reproduction, genes involved in olfactory perception have likely evolved by the action of positive natural selection. Indeed, recognition and discrimination of olfactory signals are critical for finding food sources and for the reproduction of individuals; furthermore, certain chemoreceptive processes, like pheromone perception, contribute to critical evolutionary processes such as reproductive isolation and speciation. In fact, positive natural selection has been proposed to be involved in the evolution of PBPs of the moth Chortstoneura (Lepidoptera; Willett 2000) and also in the evolution of the OBPs of fire ants and other closely related species, in which these proteins could control some aspects of social organization (Krieger and Ross 2002). Here, we analyze DNA variation at the *OS-E* and *OS-F* genes in four species of the melanogaster subgroup of Drosophila (*D. melanogaster*, *D. simulans*, *D. mauritiana*, and *D. erecta*) and also in a natural population of *D. melanogaster* to infer the evolutionary history of these genes. We found that the two genes are present in all surveyed species and thus originated from an ancient duplication event; nevertheless, these genes differ in their functional constraint. We show that the *OS-E* gene region has very distinctive evolutionary patterns, specifically, (i) an accumulation of fixed amino acid replacements in the OS-E protein of *D. melanogaster* and (ii) an atypical pattern of nucleotide polymorphism. These results suggest that positive natural selection was likely involved in the evolution of this gene.

MATERIALS AND METHODS

Fly stocks: Fourteen *D. melanogaster* isochromosomal strains for the third chromosome were used; these strains were obtained from flies collected in a natural population of Montemayor, Spain, with crosses with the TM6/MKRS balancer stock (Cirera and Aguadé 1997; Ramos-Onsins and Aguadé 1998). A highly inbred *D. simulans* line (S40; from a natural population in Montblanc, Spain), obtained by 10 generations of sib mating, was also used (Rozas *et al.* 2001). Additionally, one line of each *D. mauritiana* and *D. erecta* kindly provided by F. Lemeunier were included in the present study.

DNA extraction, PCR amplification, and DNA sequencing: Genomic DNA from the *D. melanogaster* lines was CsCl purified (Bingham *et al.* 1981). DNA from *D. simulans*, *D. mauritiana*, and *D. erecta* was extracted from a single individual by using a modification of protocol 48 in Ashburner (1989). In *D. melanogaster*, an ~4.7-kb genomic region that includes the complete coding region of both the *OS-E* and *OS-F* genes, the intergenic region, and 174 bp of the *OS-E* 5′ flanking region was amplified by PCR (Saiki *et al.* 1988; Figure 1). An additional 2-kb region upstream of the *OS-E* gene was PCR amplified in 13 of the *D. melanogaster* lines (in all lines except line M47). The amplified fragments were purified with Qiaquick columns (QIAGEN, Chatsworth, CA) and subsequently sequenced by using several oligonucleotides designed at intervals of ~400 nucleotides. The sequenced fragments were separated on ABI PRISM 377 and 3700 automated DNA sequencers. For each line, the DNA sequence was determined on both strands.

For *D. simulans* and *D. mauritiana*, the same 4.7-kb region was amplified and sequenced by using several of the primers designed for *D. melanogaster* and, for the more divergent DNA regions, by primer walking. In *D. erecta*, only the *OS-E* and *OS-F* genes were PCR amplified and sequenced. In this species, primers for amplification and sequencing were designed on the most conserved regions of the genes among the other three species and also by the primer walking technique.

Data analysis: DNA sequences were assembled using the SeqEd version 1.0.3 program (Applied Biosystems, Foster City, CA). Sequences were multiply aligned with the ClustalW program (Thompson *et al.* 1994), and the initial alignment was optimized manually. The MacClade program, version 3.06 (Maddison and Maddison 1992), was used to edit the DNA sequences for further analyses. The secondary structure of the OS-E and OS-F proteins was inferred by using the PHD and PROF secondary structure prediction programs (Rost 2001). The DNA divergence among the studied species was estimated as *K*, the number of nucleotide differences per site corrected according to Jukes and Cantor (1969). Phylogenetic analysis was performed using the neighbor-joining algorithm (Saitou and Nei 1987) implemented in the MEGA version 2 program (Kumar *et al.* 2000) and by the maximum-likelihood method (Felsenstein 1993). The bootstrap analysis was based on 1000 replicates. Coding DNA sequences at internal nodes of the phylogenetic tree were reconstructed by the maximum-likelihood ancestral reconstruction approach using codon substitution models (Goldman and Yang 1994; Yang *et al.* 1995). From the ancestral sequences, we estimated the number of synonymous and nonsynonymous substitutions in each branch. All these analyses were performed using the *codeml* program included in the PAML 3.0 software (Yang 1997).

The DnaSP version 3.98 program (Rozas and Rozas 1999) was used for most intraspecific and some interspecific analyses. The level of DNA polymorphism was estimated as the per-site nucleotide diversity (π; Nei 1987), the Watterson parameter θ (Watterson 1975), and the haplotype diversity (Hd; Nei 1987). Codon bias was measured as the effective number of codons (ENC; Wright 1990), which measures the deviation from equal usage of synonymous codons.

The recombination parameter c (in Drosophila, $c = 2Nr$, where N is the effective population size and r is the recombination rate per generation between adjacent sites) was estimated using three different methods. The Hudson (1987) method estimates c from the variance of the average number of nucleo-

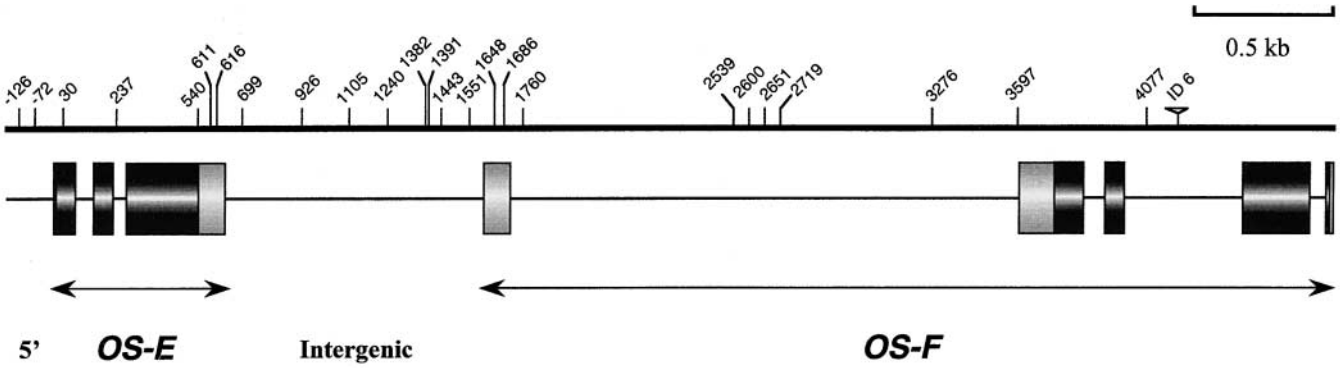

FIGURE 1.—Physical map of the *OS* region. Solid and shaded boxes indicate translated and untranslated exons, respectively. Numbers above the thick line indicate the position of the nucleotide polymorphisms detected in *D. melanogaster*. Nucleotide sites are numbered from the translation initiation site of the *OS-E* gene. The ~2-kb region upstream of the *OS-E* gene is not included in the figure.

tide differences between pairs of sequences, while the HUDSON and KAPLAN (1985) method estimates c from the minimum number of recombination events in the sample (R_M). The latter method requires the use of coalescent simulations to estimate c. An estimate of c based on the *D. melanogaster* recombination map was also obtained, assuming (i) that the recombination rate of the *OS* region (which is located on band 83CD) is 0.16×10^{-8} (see COMERON *et al.* 1999) and (ii) that N is 10^6 for *D. melanogaster*.

Statistical tests: The TAJIMA (1989) and the FU and LI (1993) tests were used to contrast whether the polymorphism frequency distribution (frequency spectrum) conforms to neutral expectations. The overall genetic association between polymorphic sites was determined by the Z_{nS} (KELLY 1997), Wall's B and Q (WALL 1999), and Z_A (ROZAS *et al.* 2001) statistics. The confidence intervals of these statistics were obtained by computer simulations (10,000 replicates) on the basis of the coalescent process assuming a large constant population size (KINGMAN 1982; HUDSON 1983, 1990; ROZAS and ROZAS 1999). Coalescent simulations were conducted using different values of the recombination parameter and conditioning on the number of segregating sites.

The Hudson-Kreitman-Aguadé (HKA) test (HUDSON *et al.* 1987) was conducted to assess whether the levels of polymorphism and divergence between species correlated, as expected under neutrality. The test was carried out using the 5′ *Adh* gene region (KREITMAN and AGUADÉ 1986) as a neutral evolving region.

The Kolmogorov-Smirnov statistic (D_{KS}) was used to test for heterogeneity in the ratio of polymorphism to divergence along the surveyed DNA region. The test is based on the maximum absolute difference between the observed and the expected cumulative number of polymorphic sites (SOKAL and ROHLF 1995; MCDONALD 1998), and it is generally the most powerful test for regions with two areas with very different levels of variation (MCDONALD 1998). The WU and LI (1985) relative-rate test was used to test for heterogeneity in the nucleotide substitution rate among lineages. This method is based on the standardized difference of the corrected estimates of the number of substitutions per site between two lineages. The K2WULI program (JERMIIN 1996) was used to perform this analysis. For nonsynonymous sites we used the χ^2 test due to the small number of substitutions.

RESULTS

Interspecific analysis: We have identified the *OS-E* and *OS-F* genes in the four species studied (*D. melanogaster*, *D. simulans*, *D. mauritiana*, and *D. erecta*). Moreover, both the intron-exon structure and the physical distance between genes are maintained across these species. Our results contrast with the previous report of HEKMAT-SCAFE *et al.* (2000), where the *OS-E* gene was not detected in either *D. simulans* or *D. mauritiana*. Nevertheless, the methodology used in their survey, a restriction-enzyme-based analysis, likely precluded its detection. Figure 2 shows the multiple alignment of the amino acid sequences encoded by the *OS-E* and *OS-F* genes. The six highly conserved cysteines of the OBP family are present in all OS proteins, except for the second cysteine of the OS-E protein in *D. erecta* that was replaced by a tryptophan. Moreover, the PHD and PROF programs predicted that all OS proteins are helical rich. To obtain clues on the function of specific parts of these proteins, the predicted structure of OS-E and OS-F was compared with that obtained for the pheromone-binding protein of *Bombix mori* (BmPBP). This protein is also a member of the OBP family and its three-dimensional (3D) structure has been determined by X-ray crystallography (SANDLER *et al.* 2000). The distribution of the predicted α-helices along the OS proteins (Figure 2) is nearly identical to that found for the BmPBP.

Nucleotide divergence between species was estimated using only the DNA sequence fragment clearly alignable among all species (Table 1); the *D. melanogaster* line M2 was used for this analysis. In general, nucleotide divergence was higher in the *OS-E* than in the *OS-F* region. Despite this difference, both genes have a very similar and quite low codon bias, with an average ENC value for all species equal to 50.55 for *OS-E* and to 45.28 for *OS-F*. In both genes, higher K_S than K_A values were detected. In the *OS-E* gene, divergence estimates were higher at synonymous than at noncoding sites.

Figure 3 shows the neighbor-joining trees reconstructed for the *OS-E* and *OS-F* genes (the same topology is obtained using the maximum-likelihood approach). In the *OS-E* tree, the branch leading to the *D. melanogaster* lineage was rather long. We conducted a relative-rate test (WU and LI 1985), using *D. erecta* as the outgroup,

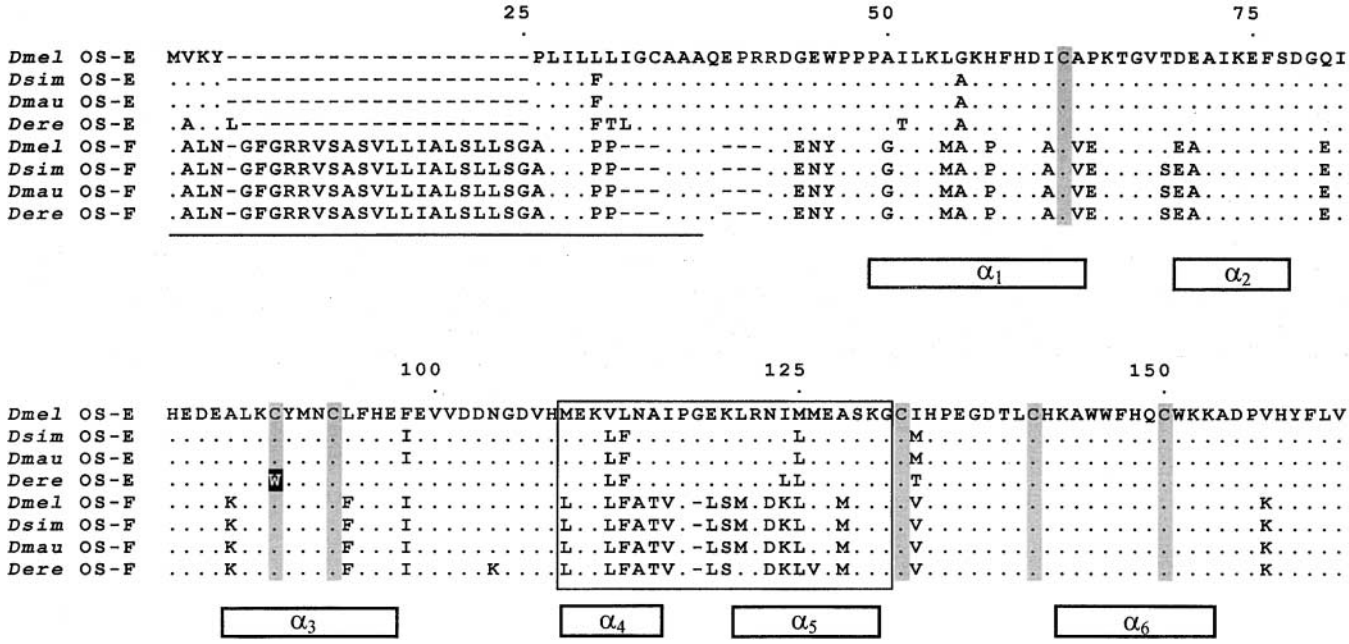

FIGURE 2.—Amino acid sequence alignment of the OS-E and OS-F proteins. Amino acids identical to the first sequence are indicated by a dot. Shaded residues indicate the six cysteines highly conserved in the OBP family. The Cys-to-Trp replacement (OS-E protein of *D. erecta*) is also indicated. The line underneath the aligned sequences indicates the predicted signal peptide. The location of the predicted α-helices is indicated by open boxes below the sequences. The box in the sequences delimits the *hr* region (see the Introduction). *Dmel, D. melanogaster*; *Dsim, D. simulans*; *Dmau, D. mauritiana*; *Dere, D. erecta.*

to determine whether the numbers of substitutions in the *D. melanogaster* and in the *D. mauritiana* (or *D. simulans*) lineages were significantly different. This test revealed that the *OS-E* region (including coding and noncoding sites) evolves faster in the *D. melanogaster* than in the *D. mauritiana* and *D. simulans* lineages ($z = 2.052$, $P = 0.021$ for *D. mauritiana*; $z = 1.395$, $P = 0.081$ for *D. simulans*). Equivalent z results were obtained when only the coding region was used ($z = 2.006$, $P = 0.024$ for *D. mauritiana*; $z = 1.994$, $P = 0.023$ for *D. simulans*). In fact, the significantly higher number of substitutions accumulated in the *D. melanogaster* lineage was mainly due to nonsynonymous substitutions ($P = 0.058$). All amino acid replacements fixed in the *D. melanogaster* lineage are conservative, *i.e.*, with very low physicochemical distance values (GRANTHAM 1974).

Nucleotide polymorphism in *D. melanogaster*: Figures 1 and 4 show the distribution of DNA polymorphic sites along the 4.7-kb region surveyed. A total of 25 nucleotide polymorphisms (9 of them with singleton variants) and 1 indel polymorphism (6 bp) were detected. All polymorphisms were silent: 1 synonymous polymorphism at site 30 of the *OS-E* coding region and the rest at noncoding positions. Notably, polymorphism at site 540 results in two different stop codons (TAG and TAA) of the *OS-E* gene. Ten different haplotypes (Hd = 0.956) were detected in the 14 lines analyzed.

Estimates of the per-site recombination parameter

TABLE 1

Nucleotide divergence in the *OS* region

Species pair	*OS-E*			*OS-F*		
	K_{NC}	K_S	K_A	K_{NC}	K_S	K_A
Dmel Dsim	0.0871	0.1102	0.0212	0.0603	0.0725	0.0028
Dmel Dmau	0.0682	0.1102	0.0212	0.0580	0.0510	0.0028
Dsim Dmau	0.0448	0.0347	0	0.0288	0.0200	0
Dmel Dere	0.2314	0.3831	0.0446	0.1359	0.1164	0.0112
Dsim Dere	0.2242	0.3514	0.0274	0.1507	0.1051	0.0084
Dmau Dere	0.2088	0.3519	0.0274	0.1367	0.0828	0.0084

K_{NC}, number of noncoding substitutions per noncoding site; K_S, number of synonymous substitutions per synonymous site; K_A, number of nonsynonymous substitutions per nonsynonymous site. *Dmel, D. melanogaster*; *Dsim, D. simulans*; *Dmau, D. mauritiana*; *Dere, D.erecta.*

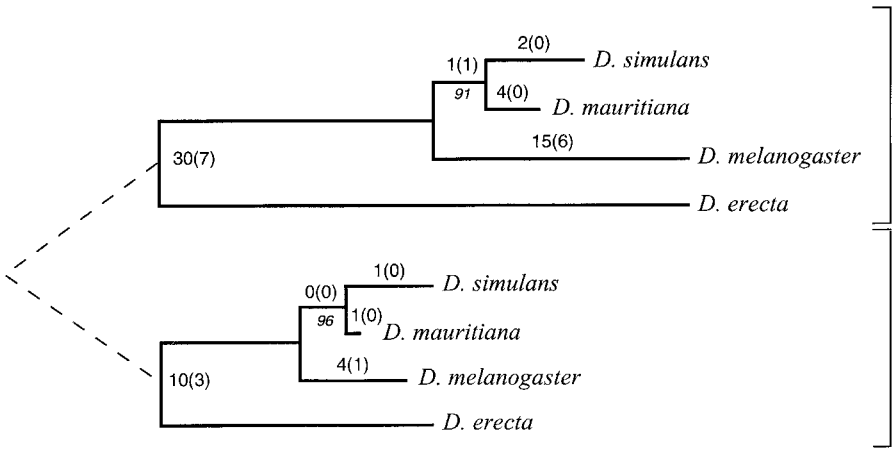

FIGURE 3.—Phylogenetic trees of the *OS-E* and *OS-F* genes. The trees were built using 651 sites (423 coding sites) and 805 sites (462 coding sites) for *OS-E* and *OS-F*, respectively. The numbers of nucleotide substitutions and the number of nonsynonymous changes (in parentheses) are indicated on each branch. The percentages of bootstrap replicates supporting the different nodes are in italics.

[c = 0.0026 from HUDSON (1987), c = 0.0024 from HUDSON and KAPLAN (1985), and c = 0.0032 from comparison of the physical and recombination maps] clearly indicate that the *OS* genomic region shows a reduction from the normal recombination levels (COMERON *et al.* 1999; ANDOLFATTO and PRZEWORSKI 2000). Table 2 summarizes the levels of nucleotide variation estimated separately for the different functional parts of the *OS* region. Estimates of nucleotide diversity for the complete 4.7-kb region (π = 0.0018, silent π = 0.0021) were rather low. However, levels of nucleotide diversity varied considerably along the *OS* region. Silent variation was highest in the *OS-E* gene (π = 0.0081) and lowest in the *OS-F* gene (π = 0.0013). Putative heterogeneity in the distribution of polymorphic to fixed silent sites along the *OS* region was tested by means of the D_{KS} test. Significant heterogeneity (P = 0.03) along the region studied was detected using the most conservative c value (see MCDONALD 1998). We also compared, by means of an HKA test, silent polymorphism and divergence in the total *OS* region (as well as separately for each functional part) and in the 5′ *Adh* region (HUDSON *et al.* 1987). A significant deviation from the neutral prediction was detected for the total *OS* region (P = 0.027). Interestingly, only variation in the *OS-F* region departed significantly from neutral predictions (*OS-E* region, P = 0.276; *OS-F* region, P = 0.017; intergenic region, P = 0.231).

We also conducted Tajima's D and Fu and Li's D and F tests separately for the three functional regions (*OS-E*, intergenic, and *OS-F*) to determine whether the frequency distribution of nucleotide variants departs from that expected under neutrality. For this analysis we used *D. mauritiana* as the outgroup. Under no recombination (which is the most conservative assumption for these tests), the analysis showed a significant deviation only in the *OS-E* region (Table 3). The significantly positive values of the Tajima's D and Fu and Li's F statistics in

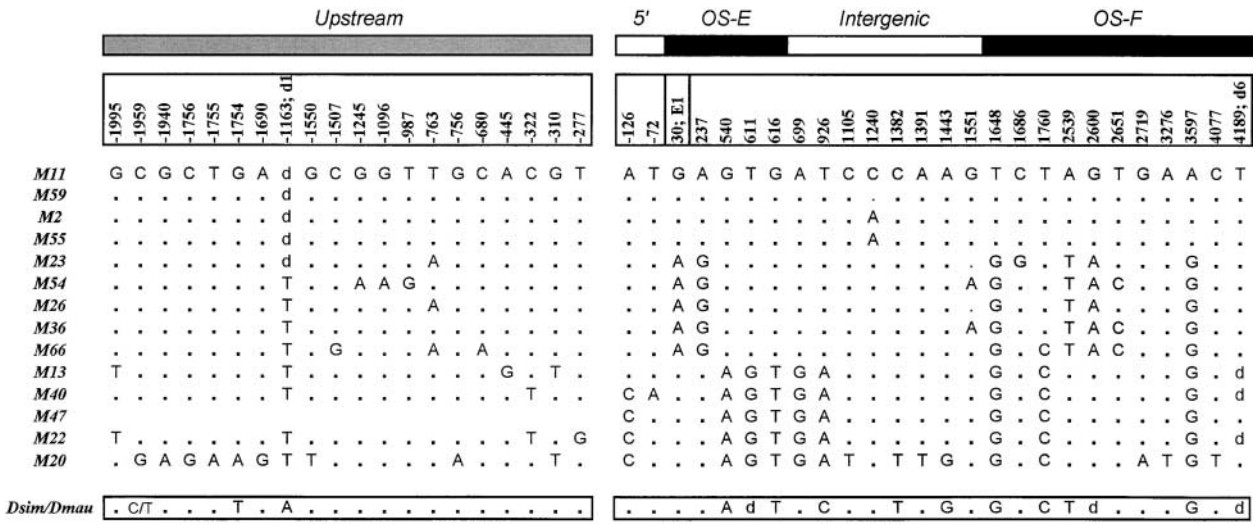

FIGURE 4.—Nucleotide polymorphisms detected in the *OS* region of *D. melanogaster*. Nucleotides identical to the first sequence are indicated by a dot. For length polymorphisms, the nucleotide position refers to the first site affected. d, deletion; d#, deletion of #bp. The last row gives, for the polymorphic positions in *D. melanogaster*, the nucleotide information in *D. mauritiana* and *D. simulans*. E1, exon 1. Information for the additional analysis of the upstream region is also shown.

TABLE 2

Nucleotide variation in different *OS* functional regions

	Sites[a]	*S*	π	θ	*K*
5′ *OS-E*	161	2	0.0036	0.0039	0.0148
OS-E					
Coding	423	1	0.0012	0.0007	0.0397
Noncoding[b]	220	4	0.0090	0.0057	0.0793
Silent[b]	307.16	5	0.0081	0.0051	0.0899
Intergenic	929	8	0.0023	0.0027	0.0505
OS-F					
Coding	462	0	0	0	0.0131
Noncoding[b]	2493	10	0.0013	0.0013	0.0470
Silent[b]	2594.33	10	0.0013	0.0012	0.0471
Total silent	3991.5	25	0.0021	0.0019	0.0498
Total	4688	25	0.0018	0.0017	0.0438

S, number of segregating sites; *K*, number of substitutions per site between *D. melanogaster* and *D. mauritiana*.
[a] Number of sites in the intraspecific data set.
[b] Including intronic and untranslated regions.

TABLE 3

Neutrality tests

	OS-E gene region	Intergenic region	*OS-F* gene region
Tajima's *D*	1.982*	0.641	0.229
Tajima's *D*[a]	1.768*	0.522	1.251
Fu and Li's *D*	1.195	0.130	−0.633
Fu and Li's *F*	1.628*	−0.227	−0.540
Z_A	0.827*	0.453	0.309
Wall's *B*	0.750*	0.428	0.222
Wall's *Q*	1.000***	0.625	0.400

*$0.01 < P < 0.05$; ***$P < 0.001$.
[a] Excluding line M20.

this region reflect an excess of nucleotide variants at intermediate frequencies. For the *OS-F* and intergenic regions, the test statistic values substantially increased when line M20 (which accounts for seven of the nine singletons found in the sample) was removed.

No overall significant association between polymorphic sites (linkage disequilibrium) was detected by the Z_{nS} statistic either in the whole region ($Z_{nS} = 0.257$; $P = 0.541$) or in its different functional parts (results not shown). Nevertheless, a significant association was detected between polymorphic sites in the *OS-E* region (4 of the 10 pairwise comparisons were significant even after the conservative Bonferroni correction). The *OS-E* region also showed significant values of the Z_A and Wall's *B* and *Q* statistics even using the conservative no-recombination assumption (Table 3). These results indicate that nucleotide variation at the *OS-E* region is highly structured (Figure 4).

DISCUSSION

The presence of both the *OS-E* and *OS-F* genes in the four Drosophila species studied, and also in species of the obscura group (A. Sánchez-Gracia, M. Aguadé and J. Rozas, unpublished results), indicates that the DNA duplication event is relatively old. Nucleotide divergence estimates among copies and among species and the phylogenetic trees clearly indicate that the two genes have evolved independently since their origin by gene duplication (*i.e.*, there is no evidence for gene conversion between paralogous copies).

Our analysis has revealed a higher number of synonymous than nonsynonymous substitutions in all phylogeny branches, suggesting that both proteins are under purifying selection. The strength of natural selection, however, differs in the two genes. Indeed, the lower nucleotide substitution rates of the *OS-F* gene indicate an overall higher functional constraint (Tables 1 and 2; Figure 3) and, therefore, would support that the idea that these genes have been functionally diverging since their origin. The detection of the two OS proteins in the *D. melanogaster* sensillar lymph (McKenna *et al.* 1994) also supports the active action of natural selection in the evolution of these genes.

It could be argued, nevertheless, that the different *OS* substitution rates were caused by local mutation rate differences and not by differential functional constraint. We found that divergence estimates at noncoding positions in the *OS-F* region are slightly lower than those in *OS-E* (Table 2). This fact is likely caused by the presence of an extremely conserved DNA fragment in the second intron of *OS-F*; this conserved region had been already identified in other species as distant as *D. virilis* (Hekmat-Scafe *et al.* 2000), although its functional significance is unknown. Levels of nucleotide divergence across noncoding regions, nevertheless, are distributed more homogeneously than in coding regions. This analysis does not support, therefore, that the different evolutionary rates found between the two *OS* genes were caused by putative differences in the silent mutation rate along the *OS* region.

Most *D. melanogaster* OBP family members (in addition to *OS-E* and *OS-F*) are located in gene clusters (Galindo and Smith 2001). This suggests that gene duplication is an important mechanism to increase diversity in this gene family. In fact, the high sequence divergence among functional members of the family suggests the contribution of positive selection to the rapid evolution and functional diversification of these genes (Galindo and Smith 2001). It has been also shown that insect OBPs can form dimers in physiological conditions

(Campanacci *et al.* 1999; Danty *et al.* 1999; Sandler *et al.* 2000). Since the *OS-E* and *OS-F* genes are coexpressed in the same cells, the encoded proteins would be able to form homodimers and also heterodimers. Although the formation of such heterodimers has not been demonstrated, this possibility is suggestive since it might be involved in the evolution of these genes: if heterodimers were more efficient than homodimers, selection might have favored the differentiation and maintenance of the two genes. Certainly, information on the quaternary structure of these OS proteins would be relevant to ascertain the role of putative dimers on the molecular evolution of these genes.

We also found an excess of substitutions at the *OS-E* coding region in the *D. melanogaster* lineage. This excess, largely due to a high number of nonsynonymous changes, could be explained by either a relaxation of natural selection or the action of positive directional selection. Although a reduction in the selective pressure could increase the fixation probability of weakly selected mutations (Ohta 1973, 1992), the reduced levels of nucleotide polymorphism in the *OS-E* coding region would not support this hypothesis. Hence, it seems likely that positive directional selection would have driven these amino acid changes to fixation (see below).

Currently, the 3D structure of the OS-E and OS-F proteins has not been determined. However, two lines of evidence suggest that this structure could be similar to that of the BmPBP obtained by X-ray crystallography. First, the secondary structures predicted for OS-E and OS-F show a remarkable similarity to that previously predicted for the BmPBP, in which the predicted location of α-helices has been confirmed by the 3D structure. Second, there are five highly conserved phenylalanines in BmPBP, with two of them (Phe12 and Phe118) involved in the general (*i.e.*, not specific) binding hydrophobic surface (Sandler *et al.* 2000). In the OS-E and OS-F proteins, there are also five conserved phenylalanines in all surveyed species. Two of these residues (Phe18 and Phe108) are also found in similar positions on the predicted BmPBP hydrophobic surface. Probably the residues constituting the odorant-binding pocket and those involved in the specific binding site of the OS-E and OS-F proteins are in locations equivalent to those described in BmPBP.

A preliminary analysis of OBPs (Plettner *et al.* 2000; Sandler *et al.* 2000; Peng and Leal 2001) has shown that some conservative amino acid changes observed across a number of Lepidoptera species might alter the protein-binding specificity. In particular, replacements among residues such as valine, leucine, isoleucine, or methionine would be responsible for this change in specificity. Remarkably, all amino acid replacements found in the four Drosophila species studied (except the Cys-to-Trp change in *D. erecta*; Figure 2) are conservative, with low physicochemical distance values (Grantham 1974), and are found in protein locations similar to the changes observed in Lepidoptera OBPs. Furthermore, four of the six OS-E amino acid changes fixed in the *D. melanogaster* lineage also involve valine, leucine, isoleucine, and methionine residues. These replacements are located either in the heterogeneous region (Hekmat-Scafe *et al.* 2000) or close to it. Some of these replacements might have been beneficial due to plausible changes in the binding specificity of the protein. Directional positive selection would thus have driven them to fixation.

The action of positive selection should have also left a fingerprint on intraspecific polymorphism and on the ratio of polymorphism to divergence. We have shown that levels of silent nucleotide polymorphism in *D. melanogaster* were reduced, which is consistent with expectations for a low-recombining genomic region (Begun and Aquadro 1992). However, the level of silent nucleotide polymorphism clearly differs between gene copies (Table 2). In fact, the observed number of segregating sites and the results of the Kolmogorof-Smirnov test ($P = 0.03$) clearly define two regions with distinct levels of variation: a left part including the *OS-E* and intergenic regions and a right part that includes the *OS-F* gene region. While the *OS-F* portion has a low level of variation, as expected in regions of reduced recombination, the level of nucleotide variation at the left part is concordant with the average level of silent variation in *D. melanogaster* (∼0.011; Moriyama and Powell 1996). The analysis of the ratio of polymorphism to divergence also shows a clear drop in about the middle of the surveyed region. The results of the HKA test are significant only in the *OS-F* region; in this part, silent nucleotide polymorphism is likely reduced due to the effects of linked selection in concordance with its low recombining environment. In contrast, no significant HKA results were obtained at the left part of the *OS* region, reflecting a much higher level of intraspecific variation than expected in a region of low recombination. In particular, there was a local peak of variation, with most polymorphisms at noncoding regions (Figure 5).

To know whether the increase of variation in *OS-E* is really a peak of variation (*i.e.*, whether it decays also in the upstream region), we sequenced an ∼2-kb region upstream of the *OS-E* gene. The level of variation in this 5′ flanking region ($\pi = 0.002$) was similar to that detected in the *OS-F* region and therefore lower than that in the *OS-E* region. The sliding window analysis of the entire region surveyed (6.7 kb) reveals a peak of variation in the *OS-E* region (Figure 5). Clearly, these results are unlikely not only under the neutral model, but also under simplistic selective models [such as the genetic hitchhiking (Maynard Smith and Haigh 1974) or the background selection (Charlesworth *et al.* 1993) models], or demographic scenarios. Heterogeneity in the recombination rate or in the silent mutation rate along the surveyed region could explain the different pattern of variation observed in the *OS-E* and *OS-F*

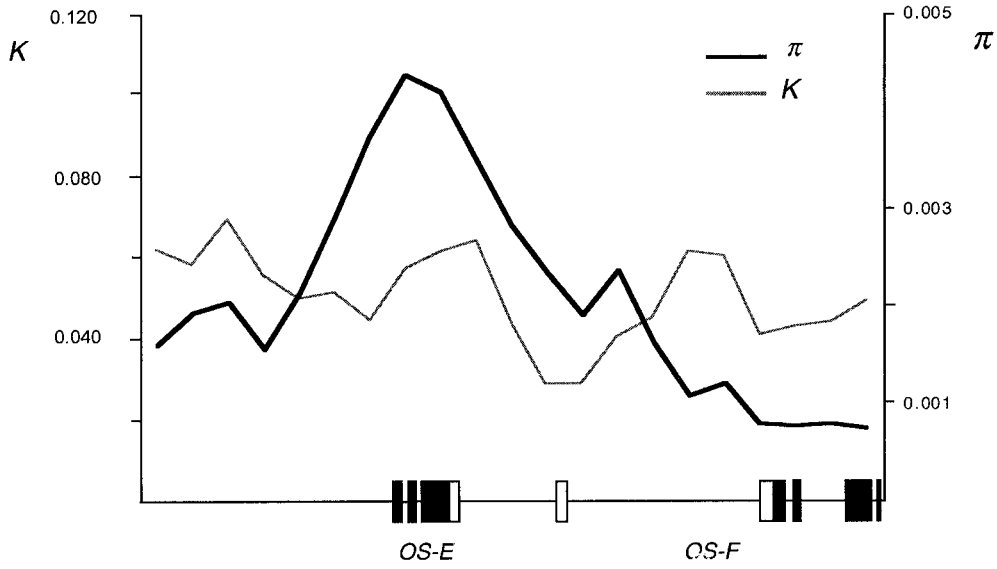

FIGURE 5.—Sliding window of silent polymorphism in *D. melanogaster* (π) and silent divergence between *D. melanogaster* and *D. mauritiana* (K) for the complete *OS* region (including the additional upstream region). Step size, 250 silent sites; window size, 1000 silent sites.

regions. However, it is unrealistic that two closely located genes could differ strongly in their recombination rate. On the other hand, the similar estimates of the silent nucleotide divergence along the *OS* region also do not support the silent mutation rate heterogeneity hypothesis.

The anomalous levels of nucleotide variation found in *OS-E* could be explained by the action of some form of balancing selection. Nevertheless, it seems difficult to envisage the target of selection since there is no replacement polymorphism. Yet, selection might act on the RNA stability or on some regulatory elements present at noncoding positions. Furthermore, several other features of the data are consistent with the balancing selection hypothesis. This kind of selection is expected to increase the levels of nucleotide variation and consequently it can skew the frequency spectrum toward intermediate frequencies. The significantly positive Tajima's *D* and Fu and Li's *F* values (Table 3) observed in the *OS-E* region are in agreement with this prediction. Nevertheless, a simple balancing selection model would not easily explain the high number of amino acid changes accumulated in the *D. melanogaster* lineage. A version of the hitchhiking model, the "traffic" model (KIRBY and STEPHAN 1996), might account for both the positive Tajima's *D* values observed in the *OS-E* gene and the increase of nucleotide diversity in the target region. This model considers that the fixation of a favorable mutation might be retarded by the fixation process of another closely located favorable mutation. Under this model, neutral variants could increase in frequency near the selected sites, reaching intermediate frequencies. Eventually, recombination could generate haplotypes with a more favorable combination of mutations that would be driven to fixation. The high structure of genetic variation in the *OS-E* region revealed by the significant linkage disequilibrium values and the Z_A and Wall's *B* and *Q* tests is consistent with the traffic hypothesis. In conclusion, the pattern of nucleotide sequence variation in the *OS* genes is unlikely under the neutral model of molecular evolution and suggests the action of positive natural selection. Further surveys of variation at genes of the olfactory family might contribute to establishing which specific mode of natural selection is acting and thereby to an understanding of the evolutionary meaning and fate of these duplicated genes.

We thank D. Hekmat-Scafe and colleagues for sending us a copy of their manuscript before publication and Serveis Científico-Tècnics, Universitat de Barcelona, for the use of automated sequencing facilities. A.S. was a predoctoral fellow of Universitat de Barcelona. This work was supported by grants PB97-0918 and BMC2001-2906 from Comisión Interdepartamental de Ciencia y Tecnología, Spain, and by grant 2001SGR-00101 from Comissió Interdepartamental de Recerca i Innovació Tecnològica, Spain, to M.A.

LITERATURE CITED

ANDOLFATTO, P., and M. PRZEWORSKI, 2000 A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics **156:** 257–268.

ASHBURNER, M., 1989 *Drosophila: A Laboratory Handbook.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

BEGUN, D. J., and C. F. AQUADRO, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *Drosophila.* Nature **356:** 519–520.

BINGHAM, P. M., R. LEVIS and G. M. RUBIN, 1981 Cloning of DNA sequences from the white locus of *D. melanogaster* by a novel and general method. Cell **25:** 693–704.

BUCK, L., and R. AXEL, 1991 A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell **65:** 175–187.

CAMPANACCI, V., S. LONGHI, P. NAGNAN-LE MEILLOUR, C. CAMBILLAU and M. TEGONI, 1999 Recombinant pheromone binding protein 1 from *Mamestra brassicae* (MbraPBP1). Eur. J. Biochem. **264:** 707–716.

CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutation on neutral molecular variation. Genetics **134:** 1289–1303.

CIRERA, S., and M. AGUADÉ, 1997 Evolutionary history of the sex-

peptide (Acp70A) gene region in *Drosophila melanogaster*. Genetics **147:** 189–197.

Comeron, J. M., M. Kreitman and M. Aguadé, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics **151:** 239–249.

Danty, E., L. Briand, C. Michard-Vanhee, V. Perez, G. Arnold *et al.*, 1999 Cloning and expression of a queen pheromone-binding protein in the honeybee: an olfactory-specific, developmentally regulated protein. J. Neurosci. **17:** 7468–7475.

Felsenstein, J., 1993 *Phylogenetic Inference Package (PHYLIP)*, Version 3.5. University of Washington, Seattle.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Galindo, K., and D. P. Smith, 2001 A large family of divergent Drosophila odorant-binding proteins expressed in gustatory and olfactory sensilla. Genetics **159:** 1059–1072.

Goldman, N., and Z. Yang, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11:** 725–736.

Grantham, R., 1974 Amino acid difference formula to help explain protein evolution. Science **185:** 862–864.

Hekmat-Scafe, D. S., R. L. Dorit and J. R. Carlson, 2000 Molecular evolution of odorant-binding protein genes *OS-E* and *OS-F* in Drosophila. Genetics **155:** 117–127.

Hekmat-Scafe, D. S., C. R. Scafe, A. J. Mckinney and M. A. Tanouye, 2002 Genome-wide analysis of the odorant-binding protein gene family in *Drosophila melanogaster*. Genome Res. **9:** 1357–1369.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

Hudson, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 245–250.

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–42 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. Futuyma and J. Antonovics. Oxford University Press, New York.

Hudson, R. R., and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

Hudson, R. R., M. Kreitman and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Jermiin, L. S., 1996 *K2WuLi*, Version 1.0. Australian National University, Canberra, Australia.

Jukes, T. H., and C. R. Cantor, 1969 *Mammalian Protein Metabolism*. Academic Press, New York.

Kelly, J., 1997 A test of neutrality based on interlocus associations. Genetics **146:** 1197–1206.

Kingman, J. F. C., 1982 On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

Kirby, D. A., and W. Stephan, 1996 Multi-locus selection and the structure of the white gene of *Drosophila melanogaster*. Genetics **144:** 635–645.

Kreitman, M., and M. Aguadé, 1986 Genetic uniformity in two populations of *Drosophila melanogaster* as revealed by filter hybridization of four-nucleotide recognizing enzyme digests. Proc. Natl. Acad. Sci. USA **83:** 3562–3566.

Krieger, M. J., and K. G. Ross, 2002 Identification of a major gene regulating complex social behavior. Science **295:** 328–332.

Kumar, S., K. Tamura, I. Jakobsen and M. Nei, 2000 *MEGA, Molecular Evolutionary Genetics Analysis*, Version 2.0.

Maddison, W. P., and D. R. Maddison, 1992 *MacClade: Analysis of Phylogeny and Character Evolution*, Version 3.0. Sinauer Associates, Sunderland, MA.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. **23:** 23–35.

McDonald, J. H., 1998 Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol. Biol. Evol. **15:** 377–384.

McKenna, M. P., D. S. Hekmat-Scafe, P. Gaines and J. R. Carlson, 1994 Putative *Drosophila* pheromone-binding proteins expressed in a subregion of the olfactory system. J. Biol. Chem. **269:** 16340–16347.

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in *Drosophila*. Mol. Biol. Evol. **13:** 261–277.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Ohta, T., 1973 Slightly deleterious substitutions in evolution. Nature **246:** 96–98.

Ohta, T., 1992 The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. **23:** 263–286.

Pelosi, P., and R. Maida, 1995 Odorant-binding proteins in insects. Comp. Biochem. Physiol. B **111:** 503–514.

Peng, G. H., and W. S. Leal, 2001 Identification and cloning of a pheromone-binding protein from the oriental beetle, *Exomala orientalis*. J. Chem. Ecol. **27:** 2183–2192.

Pikielny, C. W., G. Hasan, F. Rouyer and M. Rosbash, 1994 Members of a family of *Drosophila* putative odorant-binding proteins are expressed in different subsets of olfactory hairs. Neuron **12:** 35–49.

Plettner, E., J. Lazar, E. G. Prestwich and G. D. Prestwich, 2000 Discrimination of pheromone enantiomers by two pheromone binding proteins from the gypsy moth *Lymantria dispar*. Biochemistry **39:** 8953–8962.

Ramos-Onsins, S., and M. Aguadé, 1998 Molecular evolution of the *Cecropin* multigene family in Drosophila: functional genes *vs.* pseudogenes. Genetics **150:** 157–171.

Rost, B., 2001 Protein secondary structure prediction continues to rise. J. Struct. Biol. **134:** 204–218.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Rozas, J., M. Gullaud, G. Blandin and M. Aguadé, 2001 DNA variation at the *rp49* gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure. Genetics **158:** 1147–1155.

Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi *et al.*, 1988 Primer-directed enzymatic amplification of DNA with a thermostable polymerase. Science **239:** 487–491.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

Sandler, B. H., L. Nikonova, W. S. Leal and J. Clardy, 2000 Sexual attraction in the silkworm moth: structure of the pheromone-binding-protein-bombykol complex. Chem. Biol. **7:** 143–151.

Scaloni, A., M. Monti, S. Angeli and P. Pelosi, 1999 Structural analysis and disulfide-bridge pairing of two odorant-binding proteins from *Bombyx mori*. Biochem. Biophys. Res. Commun. **266:** 386–391.

Sokal, R., and F. J. Rohlf, 1995 *Biometry*. W. H. Freeman, New York.

Steinbrecht, R. A., 1969 Comparative morphology of olfactory receptors, pp. 3–21 in *Olfaction and Taste III*, edited by C. Pfaffmann. Rockefeller University Press, New York.

Steinbrecht, R. A., 1996 Are odorant-binding proteins involved in odorant discrimination? Chem. Senses **21:** 719–727.

Stern, D., and J. Marx, 1999 Making sense of scents. Science **286:** 703–728.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Vogt, R. G., and L. M. Riddiford, 1981 Pheromone binding and inactivation by moth antennae. Nature **293:** 161–163.

Vogt, R. G., M. E. Rogers, M. D. Franco and M. Sun, 2002 A comparative study of odorant-binding protein genes: differential expression of the PBP1-GOBP2 gene cluster in *Manduca sexta* (Lepidoptera) and the organization of OBP genes in *Drosophila melanogaster* (Diptera). J. Exp. Biol. **205:** 719–744.

Vosshall, L. B., H. Amrein, P. S. Morozov, A. Rzhetsky and R. Axel, 1999 A spatial map of olfactory receptor expression in the *Drosophila* antenna. Cell **96:** 725–736.

Vosshall, L. B., A. M. Wong and R. Axel, 2000 An olfactory sensory map in the fly brain. Cell **102:** 147–159.

Wall, J., 1999 Recombination and the power of statistical test of neutrality. Genet. Res. **74:** 65–79.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Willett, C. S., 2000 Evidence for directional selection acting on

pheromone-binding proteins in the genus *Choristoneura*. Mol. Biol. Evol. **17:** 553–562.

Wright, F., 1990 The effective number of codons used by a gene. Gene **87:** 23–29.

Wu, C.-I, and W.-H. Li, 1985 Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA **82:** 1741–1745.

Yang, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **15:** 555–556.

Yang, Z., S. Kumar and M. Nei, 1995 A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141:** 1641–1650.

Communicating editor: M. Veuille