A compact gene cluster in Drosophila: the unrelated Cs gene is compressed between duplicated amd and Ddc

Andrey Tatarenkov*, Alberto G. Sáez, Francisco J. Ayala

Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697-2325, USA

Received 16 June 1998; received in revised form 5 February 1999; accepted 10 February 1999; Received by A. Bernardi

Abstract

Cs, a gene with unknown function, and amd and Ddc, which encode decarboxylases, are among the most closely spaced genes in D. melanogaster. Untranslated 3’ ends of the convergently transcribed genes Cs and Ddc are known to overlap by 88 bp. A number of questions arise about the organization of this tightly-packed gene region and about the evolution and function of the Cs gene. We have now investigated this three-gene cluster in Scaptodrosophila lebanonensis (which diverged from D. melanogaster 60–65 MYA), as well as in D. melanogaster and D. simulans. Gene order and direction of transcription is the same in all three species. The Cs gene codes, in Scaptodrosophila, for a polypeptide of 544 amino acids; in D. melanogaster, it consists of 504 amino acids, which is twice as long as previously suggested, which makes the gene density even more spectacular. The Cs sequences exhibit higher number of non-synonymous substitutions between species, higher ratios of non-synonymous to synonymous substitutions, and lower codon usage bias than other genes, suggesting that Cs is less functionally constrained than the other genes. This is consistent with the failure of inducing phenotypic mutations in D. melanogaster. The function of Cs remains to be identified, but a high degree of similarity indicates that it is homologous to genes coding for a corticosteroid-binding protein in yeast and a polyamine oxidase in maize. © 1999 Elsevier Science B.V. All rights reserved.

Keywords: Decarboxylases, D. melanogaster; Gene cluster; Gene duplication

1. Introduction

The Ddc gene cluster in D. melanogaster, located on the left arm of the second chromosome, includes 18 identified genes plus three transcription units for which no detectable phenotypic mutations are known (Maroni, 1993; Wright, 1996; Stathakis et al., 1995). Most of the genes are densely clustered in two subclusters. Many genes in the cluster are functionally related in that they are involved in the catecholamine metabolism.

Two genes from the proximal subcluster, Ddc and amd, have been well studied, with about 90 phenotypic isolated mutations (Wright, 1996). Four genes from the proximal subcluster, including Ddc and amd, have been sequenced in D. melanogaster (Eveleth et al., 1986; Marsh et al., 1986). The coding regions of these two genes are highly similar and are thought to have arisen by gene duplication (Eveleth and Marsh, 1986). An enigmatic gene, called Cs, lies between amd and Ddc (Eveleth and Marsh, 1987). All three genes are among the most closely spaced genes in D. melanogaster, and the 3’ ends of the Ddc and Cs genes actually overlap by 88 bp (Spencer et al., 1986a; Stathakis et al., 1995). In contrast to Ddc and amd, no phenotypic mutations are known for Cs. The product of the Cs gene is not known, although its transcripts have been found associated with polysomes (Spencer et al., 1986b).

Ddc has been sequenced in a number of organisms, from mammals to insects, including D. melanogaster. Until now, the amd has been studied only in D. melanogaster, and the Cs gene is only known to occur in D. melanogaster. While Ddc and amd are members of a large family of genes, coding for PLP decarboxylases (Jackson, 1990), no genes have been reported that are similar to Cs. The origin of Cs is unknown. Its position between amd and Ddc could be a consequence of the...
original amd-Ddc duplication; or it may have been inserted there at a later time (Eveleth and Marsh, 1987).

Thus, a number of questions arise about Cs and its function and location within a developmentally important gene cluster. The first question is whether the Cs is present between amd and Ddc in other species as well, as it is in D. melanogaster. Second, the compactness of the Cs, amd, and Ddc cluster in D. melanogaster is unusual, and it is of interest to find out whether this is a result of recent events, or, rather, whether such compactness is old, perhaps tracing back to the time of the amd-Ddc duplication. One more question concerns the functional role of the Cs. As Li (1997, p. 185) has pointed out, it is well known “that the stronger the functional constraints on a macromolecule, the slower the rate of evolution.” Thus, if the Cs has a less vital function for the organism than amd and Ddc, it is expected that its evolution be faster than that of the two neighboring genes. Moreover, investigating the pattern of substitutions could help to ascertain whether the Cs is a protein encoding gene, which has been questioned (Eveleth and Marsh, 1987).

We have sequenced the Cs gene, as well as the whole amd-Cs-Ddc cluster, in the Drosophilid Scaptodrosophila lebanonensis, from a genus closely related to Drosophila. We have also sequenced in D. melanogaster Ddc and the coding region of Cs in order to resolve inconsistencies arising from previous published sequences. Finally, we have also sequenced most of the three-gene region in D. simulans for the purpose of confirming inferences about D. melanogaster.

Comparison between the Cs genes of S. lebanonensis and D. melanogaster shows high sequence similarity between them, comparable with the similarity observed for the neighboring Ddc and amd genes. Moreover, the regions of high similarity in the nucleotide and putative amino acid sequences extend much beyond the coding region previously suggested for Cs (Eveleth and Marsh, 1987). It follows that the three genes are even more tightly packed than had been previously thought for D. melanogaster, and that they are partially overlapping.

2. Materials and methods
   2.1. Species

   Isofemale lines of Drosophila melanogaster, D. simulans, and the closely related Drosophilid Scaptodrosophila lebanonensis were studied. D. melanogaster and D. simulans were collected by one of us (FJA) in St. Lucia, West Indies, in 1995. The strain of S. lebanonensis is from the National Drosophila Species Stock Center in Bowling Green, Ohio.

2.2. DNA preparation and sequencing

   Total genomic DNA was obtained using the phenol-chloroform extraction procedure described by Palumbi et al. (1991). To design amplification primers, we compared published sequences of Ddc from the moth Manduca sexta (GenBank accession number U03909), the mosquito Aedes aegypti (U27581), and D. melanogaster (X04661), as well as the amd from D. melanogaster (X04695). Ddc and amd in D. melanogaster are quite similar to each other in sequence but have different orientation. We selected segments of the aligned sequences, that had high similarity but also specific substitutions in the amd sequence when compared with Ddc sequences. The two primers (forward 5'-GAYATYGARCGNGTSATCATGCCKGG-3', and reverse 5'-GAYATYAGCGNGTSATCAAGCCG-3') encompass large parts of Ddc and amd as well as the interval between them (Fig. 1). A region of about 5.8 kb was obtained in several species of Drosophilidae.

   PCR reactions were performed in a 100 μl volume of the ExTAKARA buffer containing 2.5 U of ExTAKARA Taq polymerase, 0.5 μM each of the forward and reverse primers, 0.2 mM dNTP, and 3 μl of genomic DNA. The cycling parameters for the amplification were an initial denaturation at 95°C for 5 min and 31 cycles of the following: denaturation for 30 s at 95°C, annealing for 1 min at 60°C, and extension for 3 min at 72°C for the first cycle and an extra 3 s for every subsequent cycle; after 31 cycles the reaction was additionally kept at 72°C for 7 min to complete extension.

   The PCR product of S. lebanonensis was purified with Wizard PCR prep DNA purification system (Promega Corporation), and cloned using the TA cloning kit (Invitrogen, San Diego, CA). DNA sequencing was partly done by the dideoxy chain-termination technique with Sequenase Version 2.0 T7 DNA polymerase (Amersham Life Sciences Inc., USA) using 32P-labeled dATP, and partly with an ABI model 373 autossequencer using Dye Terminator Ready Reaction Kit in accordance with the manufacture protocol (Perkin Elmer) (see Fig. 1). We employed a successive approach for sequencing the region, so that new sequencing primers were designed based on the sequence obtained with previous primers. Both strands were completely sequenced with 34 primers.

   Sequences of the Cs gene in both D. melanogaster and D. simulans were obtained by direct sequencing of purified PCR products with an ABI model 377 autossequencer using the Dye Terminator Ready Reaction Kit in accordance with the manufacturer’s protocol (Perkin Elmer). Partial sequences of Ddc in D. melanogaster and D. simulans were obtained from separately constructed clones of these species. The sequences of these clones overlap considerably with the PCR frag-
A DNA fragment of approximately 5.8 kb resulted from PCR amplification in several drosophilid species, S. lebanonensis, D. melanogaster, D. simulans, D. immigrans, D. mimica, D. (Scaptomyza) pallae, and D. (Samosia) leonensis. The gene organization of the amplified region in D. melanogaster and S. lebanonensis is outlined in Fig. 1.

We searched the region between the stop codons of Ddc and amd in S. lebanonensis, presumably corresponding to the Cs gene, seeking segments similar with the sequence of Cs in D. melanogaster (X05991). We found an extended region, about 1.5 kb with high similarity (71%) to the sequence of Cs in D. melanogaster (Figs. 1 and 2). Unexpectedly, the region of similarity extends more than 400 bp beyond the previously suggested Cs stop codon in D. melanogaster (Eveleth and Marsh, 1986). The analysis was kindly conducted by R.C. Lewontin. Rates of substitution at synonymous and non-synonymous sites were calculated by the method of Li (1993). We searched GenBank sequences with the BLAST at http://www.ncbi.nlm.nih.gov/.

3. Results

A DNA fragment of approximately 5.8 kb resulted from PCR amplification in several drosophilid species, S. lebanonensis, D. melanogaster, D. simulans, D. immigrans, D. mimica, D. (Scaptomyza) pallae, and D. (Samosia) leonensis. The gene organization of the amplified region in D. melanogaster and S. lebanonensis is outlined in Fig. 1.

We searched the region between the stop codons of Ddc and amd in S. lebanonensis, presumably corresponding to the Cs gene, seeking segments similar with the sequence of Cs in D. melanogaster (X05991). We found an extended region, about 1.5 kb with high similarity (71%) to the sequence of Cs in D. melanogaster (Figs. 1 and 2). Unexpectedly, the region of similarity extends more than 400 bp beyond the previously suggested Cs stop codon in D. melanogaster (Eveleth and Marsh, 1986). The analysis was kindly conducted by R.C. Lewontin. Rates of substitution at synonymous and non-synonymous sites were calculated by the method of Li (1993). We searched GenBank sequences with the BLAST at http://www.ncbi.nlm.nih.gov/.
Fig. 2. Alignment of the Cs coding region between D. melanogaster (MEL), D. simulans (SIM), and S. lebanonensis (LEB). The intron in D. melanogaster is shown in lowercase letters at the top of the figure; the proposed initiator ATGs are underlined; the stop codons are in bold. A region of uncertain alignment is overscored with a double dotted line (top). Dots indicate nucleotides identical to D. melanogaster; hyphens indicate gaps. Discrepancies between our Cs sequence and the sequence of D. melanogaster of Eveleth and Marsh (1987) are shown with rectangles to indicate nucleotides missing in their sequence, and arrows to indicate locations at which they show excessive number of nucleotides.
Moreover, the *S. lebanonensis* sequence has high similarity to a segment upstream of the largest ORF previously identified (Eveleth and Marsh, 1987) in the *D. melanogaster* Cs gene. This whole 1.5 kb region is an uninterrupted open reading frame (ORF) in *S. lebanonensis*. While the *S. lebanonensis* and *D. melanogaster* sequences are highly similar at the nucleotide level along the whole 1.5 kb region, the corresponding peptide sequences are similar only in a few stretches, which are interrupted by stretches that cannot be aligned. This appears to be a consequence of shifts in reading frame due to indels in the published sequence of *D. melanogaster* (Eveleth and Marsh, 1987) compared with *S. lebanonensis*.

In order to test these inferences, we sequenced the Cs gene and adjacent regions in *D. melanogaster*, as well as in the closely related *D. simulans*. Our Cs sequence of *D. melanogaster* differs from the published sequence by the occurrence of nine indels, as predicted by the alignment of the previously published sequence with the Cs sequence of *S. lebanonensis* (see Fig. 2).

The corrected sequence of Cs in *D. melanogaster* is very similar to the Cs sequence of *D. simulans*. In both species we found a long ORF that extends for 1507 bp from the intron, determined in *D. melanogaster* by comparison of our genomic sequence with the cDNA sequence of Eveleth and Marsh (1987). The longest ORF previously proposed is 735 bp. Thus, the coding region of Cs is twice as long as previously thought (Eveleth and Marsh, 1987). In addition, the putative amino acid sequence differs from the one previously suggested for *D. melanogaster* in several stretches, some as long as 30 amino acids. The Cs stop codons of the three species are in corresponding positions on our aligned sequences, although several gaps are necessary in order to obtain the alignment (Fig. 2). The alignment of the encoded peptide sequences obtained by translating the ORF yields 95% amino acid identity between *D. melanogaster* and *D. simulans*, and 78% between *Scaptodrosophila* and those two species.

Although the similarity of the inferred coding regions is high, this high similarity does not start from the very beginning of the coding region. We are thus unable to use sequence comparisons between *D. melanogaster* and *S. lebanonensis* for elucidating the whole length of the coding regions. This is not surprising because the segment of the first exon in *D. melanogaster* is very short, just three codons, according to Eveleth and Marsh (1987). We have used several methods to infer the start of the coding region in *S. lebanonensis*, and have applied the same methods also to *D. melanogaster*. The programs GENIE and FGENED both predict an intron on the *D. melanogaster* sequence as detected by Eveleth and Marsh (1987) by comparing cDNA with genomic DNA. They also predict the first short exon postulated by Eveleth and Marsh (1987). The first eight nucleotides of the first exon merge with the remaining 1507 bp of the long ORF that we have found in *D. melanogaster*. FGENED suggests that the coding region of Cs in *S. lebanonensis* consists of just a single exon, which starts 21 codons upstream of the region where similarity between *D. melanogaster* and *S. lebanonensis* can be detected. GENIE yields the same start and stop codons as FGENED. However, GENIE indicates the presence in *Scaptodrosophila* of a short intron (positions 587–718 in Fig. 2). We rather assume that this is a coding segment given that it is highly similar to the sequences of *D. melanogaster* and *D. simulans* along the segment’s whole length at both the nucleotide and the amino acid level. It is also possible that the Cs in *D. melanogaster* consisted of a continuous single exon in the past, and that an intron (62 bp) may have arisen due to mutations that have disrupted the beginning of the coding sequence. This would explain the somewhat unusual position of the intron, after an exon of only three codons. It is also possible, but seems less likely, that an intron in the ancestral species may have become a coding sequence in *S. lebanonensis* as a result of mutation in the intron’s splice site. The predicted peptide length of Cs in *D. melanogaster* is 504 amino acids, compared with 544 amino acids in *S. lebanonensis*, if we assume a single exon.

The regions suggested as protein coding regions are characterized by somewhat increased codon bias along their length (not shown), which is indicative of coding regions (Gribkov et al., 1984). Fig. 3 shows the effective number of codons, ENC, for six genes, including Cs and the flanking *amd* and *Ddc* genes, in the three species, *S. lebanonensis*, *D. melanogaster*, and *D. simulans*. In all three species, codon-use is less biased for Cs than for any of the other genes, although it is rather similar to that for *amd* (ENC = 61 when all codons are evenly used, ENC = 20 when only one codon per amino acid is used).

### 4. Discussion

*amd*, Cs, and Ddc are neighboring genes in *D. melanogaster* (Eveleth and Marsh, 1986). *amd* and Ddc are quite similar in nucleotide and amino acid sequences, and are paralogous genes arising from an ancient gene duplication (Eveleth and Marsh, 1986; Wang et al., 1996). Ddc has been sequenced in a number of organisms (Tatarenkov et al., 1999), but the *amd* and Cs sequences have been reported only for *D. melanogaster*. Comparison of the Ddc sequences available in GenBank with those of *amd* from a number of species (our unpublished data) suggests that the duplication of these genes occurred well before the split of Lepidoptera and Diptera and may predate the divergence of Protostoma and Deuterostoma, which occurred more than 600
MYA, before the Cambrian (Jackson, 1990). If this inference is correct, amd should be present in many animal phyla, unless it has been obliterated, or has evolved beyond recognition. The physical proximity of amd and Ddc most likely traces back to the time of the original duplication of these genes, but the presence of Cs between them is enigmatic. It could reflect the survival by a gene that was contiguous to the duplicated gene that led to Ddc and amd, or it may have been inserted there at a much later time. The position in diverse animal groups of the orthologous genes to these three might permit us to resolve this issue.

We have amplified and sequenced a 5.8 kb-long fragment of genomic DNA comprising partially the flanking Ddc and amd genes and an intermediate region in D. melanogaster, D. simulans, and S. lebanonensis, a drosophilid species which diverged from D. melanogaster about 60–65 MYA (Kwiatowski et al., 1994, 1997). A PCR fragment of similar length was obtained from several other Drosophilids. The fact that the region has remained unchanged in several independent lineages during the last 30–40 Myr may be indication of its functional importance. The comparison of the region between amd and Ddc in S. lebanonensis and D. melanogaster has revealed the presence of the Cs gene in S. lebanonensis, as it was already known in D. melanogaster (Eveleth and Marsh, 1987). Moreover, this Cs gene is also present in D. simulans, where the sequence and exon-intron arrangement is extremely similar to our sequence of D. melanogaster (but importantly different at a few nucleotide sites from a previously published sequence; see Eveleth and Marsh, 1987). However, it is still not possible to answer when Cs arose between amd and Ddc. Comparisons with species distantly related to Drosophila are necessary, such as remote dipterans, other insects, and Crustacea. These comparisons will also help in dating the time of the ancestral duplication leading to amd and Ddc. Cs codes for a product of 544 amino acids in D. melanogaster, but 504 amino acids in D. simulans. The larger size than previously proposed of the postulated coding region in D. melanogaster is robust, because in addition to such characteristics of coding regions as increased GC bias and certain codon preferences, the predicted polypeptides exhibit high sequence similarity (95% between D. melanogaster and D. simulans, 75% between them and S. lebanonensis), which would be unexpected in non-coding regions.

The great proximity of the three genes, amd, Cs and Ddc in D. melanogaster is quite unusual (see discussion about 60–65 MYA (Kwiatowski et al., 1994, 1997)). A PCR fragment of similar length was obtained from several other Drosophilids. The fact that the region has remained unchanged in several independent lineages during the last 30–40 Myr may be indication of its functional importance. The comparison of the region between amd and Ddc in S. lebanonensis and D. melanogaster has revealed the presence of the Cs gene in S. lebanonensis, as it was already known in D. melanogaster (Eveleth and Marsh, 1987). Moreover, this Cs gene is also present in D. simulans, where the sequence and exon-intron arrangement is extremely similar to our sequence of D. melanogaster (but importantly different at a few nucleotide sites from a previously published sequence; see Eveleth and Marsh, 1987). However, it is still not possible to answer when Cs arose between amd and Ddc. Comparisons with species distantly related to Drosophila are necessary, such as remote dipterans, other insects, and Crustacea. These comparisons will also help in dating the time of the ancestral duplication leading to amd and Ddc. Cs codes for a product of 544 amino acids in D. melanogaster, but 504 amino acids in D. simulans. The larger size than previously proposed of the postulated coding region in D. melanogaster is robust, because in addition to such characteristics of coding regions as increased GC bias and certain codon preferences, the predicted polypeptides exhibit high sequence similarity (95% between D. melanogaster and D. simulans, 75% between them and S. lebanonensis), which would be unexpected in non-coding regions.

The great proximity of the three genes, amd, Cs and Ddc in D. melanogaster is quite unusual (see discussion about 60–65 MYA (Kwiatowski et al., 1994, 1997)). A PCR fragment of similar length was obtained from several other Drosophilids. The fact that the region has remained unchanged in several independent lineages during the last 30–40 Myr may be indication of its functional importance. The comparison of the region between amd and Ddc in S. lebanonensis and D. melanogaster has revealed the presence of the Cs gene in S. lebanonensis, as it was already known in D. melanogaster (Eveleth and Marsh, 1987). Moreover, this Cs gene is also present in D. simulans, where the sequence and exon-intron arrangement is extremely similar to our sequence of D. melanogaster (but importantly different at a few nucleotide sites from a previously published sequence; see Eveleth and Marsh, 1987). However, it is still not possible to answer when Cs arose between amd and Ddc. Comparisons with species distantly related to Drosophila are necessary, such as remote dipterans, other insects, and Crustacea. These comparisons will also help in dating the time of the ancestral duplication leading to amd and Ddc. Cs codes for a product of 544 amino acids in D. melanogaster, but 504 amino acids in D. simulans. The larger size than previously proposed of the postulated coding region in D. melanogaster is robust, because in addition to such characteristics of coding regions as increased GC bias and certain codon preferences, the predicted polypeptides exhibit high sequence similarity (95% between D. melanogaster and D. simulans, 75% between them and S. lebanonensis), which would be unexpected in non-coding regions.

The great proximity of the three genes, amd, Cs and Ddc in D. melanogaster is quite unusual (see discussion about 60–65 MYA (Kwiatowski et al., 1994, 1997)). A PCR fragment of similar length was obtained from several other Drosophilids. The fact that the region has remained unchanged in several independent lineages during the last 30–40 Myr may be indication of its functional importance. The comparison of the region between amd and Ddc in S. lebanonensis and D. melanogaster has revealed the presence of the Cs gene in S. lebanonensis, as it was already known in D. melanogaster (Eveleth and Marsh, 1987). Moreover, this Cs gene is also present in D. simulans, where the sequence and exon-intron arrangement is extremely similar to our sequence of D. melanogaster (but importantly different at a few nucleotide sites from a previously published sequence; see Eveleth and Marsh, 1987). However, it is still not possible to answer when Cs arose between amd and Ddc. Comparisons with species distantly related to Drosophila are necessary, such as remote dipterans, other insects, and Crustacea. These comparisons will also help in dating the time of the ancestral duplication leading to amd and Ddc. Cs codes for a product of 544 amino acids in D. melanogaster, but 504 amino acids in D. simulans. The larger size than previously proposed of the postulated coding region in D. melanogaster is robust, because in addition to such characteristics of coding regions as increased GC bias and certain codon preferences, the predicted polypeptides exhibit high sequence similarity (95% between D. melanogaster and D. simulans, 75% between them and S. lebanonensis), which would be unexpected in non-coding regions.

The great proximity of the three genes, amd, Cs and Ddc in D. melanogaster is quite unusual (see discussion about 60–65 MYA (Kwiatowski et al., 1994, 1997)). A PCR fragment of similar length was obtained from several other Drosophilids. The fact that the region has remained unchanged in several independent lineages during the last 30–40 Myr may be indication of its functional importance. The comparison of the region between amd and Ddc in S. lebanonensis and D. melanogaster has revealed the presence of the Cs gene in S. lebanonensis, as it was already known in D. melanogaster (Eveleth and Marsh, 1987). Moreover, this Cs gene is also present in D. simulans, where the sequence and exon-intron arrangement is extremely similar to our sequence of D. melanogaster (but importantly different at a few nucleotide sites from a previously published sequence; see Eveleth and Marsh, 1987). However, it is still not possible to answer when Cs arose between amd and Ddc. Comparisons with species distantly related to Drosophila are necessary, such as remote dipterans, other insects, and Crustacea. These comparisons will also help in dating the time of the ancestral duplication leading to amd and Ddc. Cs codes for a product of 544 amino acids in D. melanogaster, but 504 amino acids in D. simulans. The larger size than previously proposed of the postulated coding region in D. melanogaster is robust, because in addition to such characteristics of coding regions as increased GC bias and certain codon preferences, the predicted polypeptides exhibit high sequence similarity (95% between D. melanogaster and D. simulans, 75% between them and S. lebanonensis), which would be unexpected in non-coding regions.
Table 1

Number of non-synonymous (n-syn) and synonymous (syn) substitutions per site ± SE, and their ratio (n-syn/syn), between *Drosophila melanogaster*, *D. simulans*, and *Scaptodrosophila lebanonensis* at six nuclear genes. The sequences of *Adh* are from Russo et al. (1995); *Gpdh* from Kwiatowski et al. (1997); *Sod* from Kwiatowski et al. (1994).

<table>
<thead>
<tr>
<th></th>
<th>melanogaster–simulans</th>
<th>melanogaster–lebanonensis</th>
<th>simulans–lebanonensis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n-syn</td>
<td>syn</td>
<td>ratio</td>
</tr>
<tr>
<td><em>Amd</em></td>
<td>0.010 ± 0.004</td>
<td>0.105 ± 0.013</td>
<td>0.099 ± 0.012</td>
</tr>
<tr>
<td></td>
<td>0.151 ± 0.027</td>
<td>1.315 ± 0.174</td>
<td>1.372 ± 0.188</td>
</tr>
<tr>
<td></td>
<td>0.066</td>
<td>0.080</td>
<td>0.072</td>
</tr>
<tr>
<td><em>Cs</em></td>
<td>0.027 ± 0.005</td>
<td>0.212 ± 0.016</td>
<td>0.217 ± 0.016</td>
</tr>
<tr>
<td></td>
<td>0.149 ± 0.022</td>
<td>1.562 ± 0.204</td>
<td>1.467 ± 0.186</td>
</tr>
<tr>
<td></td>
<td>0.181</td>
<td>0.136</td>
<td>0.148</td>
</tr>
<tr>
<td><em>Ddc</em></td>
<td>0.003 ± 0.002</td>
<td>0.066 ± 0.010</td>
<td>0.064 ± 0.010</td>
</tr>
<tr>
<td></td>
<td>0.064 ± 0.010</td>
<td>1.239 ± 0.175</td>
<td>1.155 ± 0.156</td>
</tr>
<tr>
<td></td>
<td>0.047</td>
<td>0.053</td>
<td>0.055</td>
</tr>
<tr>
<td><em>Adh</em></td>
<td>0.002 ± 0.002</td>
<td>0.101 ± 0.016</td>
<td>0.103 ± 0.017</td>
</tr>
<tr>
<td></td>
<td>0.052 ± 0.021</td>
<td>0.802 ± 0.122</td>
<td>0.765 ± 0.117</td>
</tr>
<tr>
<td></td>
<td>0.038</td>
<td>0.126</td>
<td>0.135</td>
</tr>
<tr>
<td></td>
<td>0.000 ± 0.000</td>
<td>0.032 ± 0.005</td>
<td>0.031 ± 0.005</td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.053</td>
<td>0.055</td>
</tr>
<tr>
<td><em>Gpdh</em></td>
<td>0.000 ± 0.000</td>
<td>0.113 ± 0.020</td>
<td>0.106 ± 0.019</td>
</tr>
<tr>
<td></td>
<td>0.114 ± 0.037</td>
<td>1.508 ± 0.473</td>
<td>1.938 ± 1.030</td>
</tr>
<tr>
<td></td>
<td>0.000</td>
<td>0.075</td>
<td>0.056</td>
</tr>
</tbody>
</table>

for a considerable time, at least 60–65 Myr in *S. lebanonensis* and *D. melanogaster*. The structure and sequence of this region have remained essentially identical in *D. simulans* and *D. melanogaster*, that is for some 2.5 Myr (we have not investigated the region upstream of the *Cs* coding sequence in *D. simulans*, but it also seems quite similar with respect to length, since the PCR fragments are of similar length). Eveleth and Marsh (1987) failed to recover *Cs* phenotypic mutants in their extensive mutagenesis screens and suggested that this implies that the *Cs* function is not essential or that *Cs* RNA does not encode a protein, as

Fig. 4. Fourier transform of the substitution pattern in four genes: *amd, Ddc, Cs*, and *Hsr-omega* are from *D. melanogaster*, *D. simulans*, and *S. lebanonensis*, *Hsr-omega* is from *D. melanogaster* and *D. pseudoobscura*. A dominant substitution frequency of 1/3 is revealed for the exon sequences of *amd*, *Ddc*, and *Cs*, while no predominant peak is observed for the two exons of the non-coding *hhr-omega* (a peak very close to the y-axis is due to non-specific correlations and is largely diminished when gaps are eliminated from the alignment; data not shown).
described for hsr-omega in Drosophila (Fini et al., 1989). We propose, however, that Cs retains protein-encoding capacity. Thus, the ratio of non-synonymous to synonymous substitutions between D. melanogaster and D. simulans, or between the later two and S. lebanonensis, is much less than 1 (Table 1). In a complementary analysis (which may be more appropriate for a comparison between distantly related species, such as S. lebanonensis with respect to D. melanogaster/D. simulans, because of the possible saturation at synonymous sites) the pattern of substitutions also indicates a protein-encoding capacity for Cs. If so, Cs should have a dominant peak of periodical substitutions every third base. Periodicity in DNA sequences can be unveiled using Fourier analysis (Tsonis et al., 1991), which we have investigated using an algorithm by Cooley and Tukey (1965). A clear dominant substitution frequency of 1/3 (i.e. every third position) is observed for amd, Ddc, but also for Cs, while it is not for hsr-omega (Fig. 4), although the percentage of substitutions is similar in the four genes: 27.8%, 21.6%, 36.6%, and 24.4%, respectively. Nevertheless, there is some indication that the selective constraints may be somewhat lower for Cs than for other genes. Thus, the ratio of non-synonymous to synonymous substitutions is higher in Cs than in the other five genes (Table 1). Additionally, the number of non-synonymous substitutions per site is higher in Cs. Moreover, several gaps, some as long as 10 codons, are needed to align the Cs sequences of the three studied species, whereas only one or three gaps are required in amd and Ddc.

The hypothesis of lesser functional constraints imposed on Cs is furthermore supported by analysis of codon usage bias, which is lowest in Cs for all three Drosophilidae that we have studied (Fig. 3). Irrespective of the mechanism underlying the natural selection on silent sites (e.g. rates of protein elongation, translational accuracy), codon usage is typically most biased in highly expressed genes with high functional constraints (Shields et al., 1988; Moriyama and Hartl, 1993; Akashi, 1994; Moriyama and Powell, 1997). Note, however, that although the codon usage bias in Cs is not as pronounced as in such highly expressed genes as Adh and Sod, it is not atypical for Drosophila. Particularly, ENC in Cs is rather close to that in the neighboring amd. Earlier observations that codon usage bias in Cs is very weak compared with other D. melanogaster genes (Eveleth and Marsh, 1987; Statilaks et al., 1995) may have arisen from mistakes in the previously published sequence of Cs.

We have studied the spatial distribution of substitutions in the deduced amino acid sequences of amd, Cs, and Ddc (Fig. 5). The three proteins show seemingly different distribution of the substitutions, with Cs appearing as the most homogeneous. However, the unmodified variance test of Goss and Lewontin (1996) reveals statistically significant non-random clustering of substitutions ($P<0.01$) in all three genes in the comparison between D. melanogaster and S. lebanonensis, both including and excluding the conservative substitutions (short lines in Fig. 5). Interestingly, the non-randomness is more pronounced for all three proteins when conservative substitutions are not considered (i.e. $P$ values are smaller). A graphical plot of the distribution of the segment sizes between substitutions shows an excess,
As shown above, Cs appears to be a protein-encoding gene. Consequently, we have conducted an extensive search of GenBank for sequences that would be similar to Cs, and have at least six sequences that are distantly related to Cs, although more similar than expected by chance. Fig. 6 displays the protein alignments with the two most similar sequences: a corticosteroid-binding protein in the yeast Candida albicans, and a polyamine oxidase in maize, Zea mays. Although the similarity of these sequences to Cs is not very high, they are surely homologous. First, the probability of the sequence similarity observed is in both cases \( P < 10^{-6} \). Moreover, the alignment encompasses large segments of the genes: about 90% of the Cs gene in S. lebanonensis and 95% in D. melanogaster and virtually the whole extension of the coding regions in the genes of Candida and Zea. Other sequences of about the same length and with similarity nearly as large include amine oxidase in a fish (P49253), protoporphyrinogen oxidase in tobacco (Y13466), and proteins of unknown function with similarities to monoamine oxidase and protein kinase in Caenorhabditis (z78198, locus 1491653) and Arabidopsis (G2244987).

In S. lebanonensis the protein is encoded by a single ORF, while in D. melanogaster the coding sequence is interrupted by a short intron.

Fig. 6. Similarity between the deduced Cs protein in S. lebanonensis and two other proteins: corticosteroid-binding protein in the yeast Candida albicans (PIR: A47259), and polyamine oxidase in maize (GenBank: AJ002204). The numbers at the two ends of each row refer to amino acid sites in the proteins. Identical amino acids are shown by letters in the middle rows; crosses indicate functionally similar amino acids.

### Table: Corticosteroid-binding protein in yeast (Candida albicans) and Polyamine oxidase in maize (Zea mays)

<table>
<thead>
<tr>
<th>Corticosteroid-binding protein</th>
<th>Polyamine oxidase in maize</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>S. lebanonensis</strong></td>
<td><strong>D. melanogaster</strong></td>
</tr>
<tr>
<td>yest 17</td>
<td>yest 63</td>
</tr>
<tr>
<td>yest 17</td>
<td>yest 63</td>
</tr>
<tr>
<td>yest 22</td>
<td>yest 24</td>
</tr>
<tr>
<td>yest 23</td>
<td>yest 25</td>
</tr>
<tr>
<td>yest 33</td>
<td>yest 36</td>
</tr>
<tr>
<td>yest 40</td>
<td>yest 45</td>
</tr>
<tr>
<td>yest 44</td>
<td>yest 50</td>
</tr>
<tr>
<td>yest 50</td>
<td>yest 53</td>
</tr>
<tr>
<td>yest 52</td>
<td>yest 55</td>
</tr>
</tbody>
</table>

5. Conclusions

(1) Gene order and direction of transcription of the amd, Cs, and Ddc genes are the same in S. lebanonensis and D. melanogaster. Cs is very closely packed with the neighboring Ddc and amd genes in S. lebanonensis as well as in Drosophila.

(2) The Cs gene codes for a longer product than had been previously suggested for D. melanogaster. The length of the deduced protein is 544 amino acids in S. lebanonensis and 504 amino acids in D. melanogaster.

In S. lebanonensis the gene is encoded by a single ORF, while in D. melanogaster the coding sequence is interrupted by a short intron.

(3) There is heterogeneity in substitution pattern between and within amd, Cs, and Ddc. Cs appears to be the most constrained gene of the three, especially its central area. amd is less constrained, with a highly variable carboxyl end and a more conserved central area.

(4) Compared with some other nuclear genes, the Drosophilidae Cs sequences exhibit higher number of non-synonymous substitutions, higher ratios of non-synonymous to synonymous substitutions, and lower
codon usage bias, suggesting that Cr is not functionally so highly constrained as the other genes.

(5) The Cr protein exhibits statistically significant sequence similarity to other proteins, such as some oxidases.

Acknowledgements

We thank R.C. Lewontin for running the unmodified variance test for us; Shiliang Qin and John W. Jacobs for use of an automatic sequencer of the Hitachi Chemical Research Center, Inc. at the University of California, Irvine; and José L. Oliver, Wentian Li, and Hubert Berens for help with the Fourier analysis. Research supported by NIH Grant GM42397 to F.J.A.

References


Wisconsin Package Version 9.1, Genetics Computer Group. Madison, WI.
