ELSEVIER

Gene 231 (1999) 111–120

GENE
AN INTERNATIONAL JOURNAL ON GENES AND GENOMES

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 92697-2525, 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.

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

Abbreviations: *amd*, *α-methyl dopa sensitive* gene encoding decarboxylase related enzyme (product unknown); bp, base pair(s); BLAST, basic local alignment search tool; *Cs*, a gene with unknown function; *Ddc*, gene encoding Dopa decarboxylase (DDC, EC 4.1.1.26); ENC, effective number of codons; Myr, million years; MYA, million years ago; PCR, polymerase chain reaction; PLP, pyridoxal 5′-phosphate.

* Corresponding author. Fax: +1-949-824-2474.

E-mail address: antatare@uci.edu (Andrey Tatarenkov)

0378-1119/99/$ – see front matter
PII: S0378-1119(99)00096-7

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′-GAYATYAGYCGNGTSATCAAGCCKGG-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 5 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 preps 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 ^{35}S-labeled dATP, and partly with an ABI model 373 autosequencer 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 autosequencer 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-

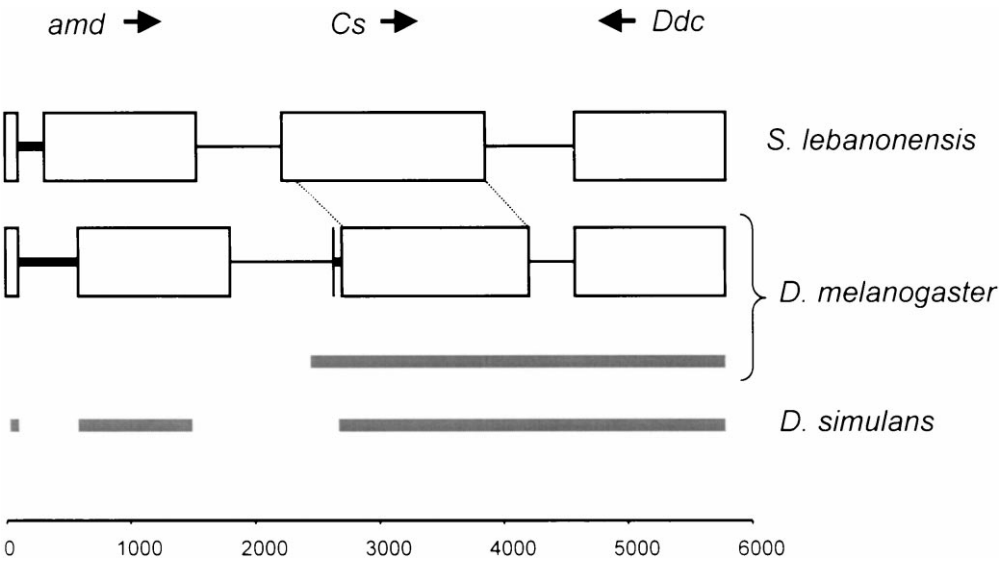

Fig. 1. Structure, gene arrangement, and direction of transcription of a genomic DNA segment comprising the genes *amd*, *Cs*, and *Ddc* in *Scaptodrosophila lebanonensis* and *Drosophila melanogaster*. Thick arrows adjacent to gene symbols indicate direction of transcription from 5′ to 3′. Boxes indicate protein coding regions: thick lines connecting them represent introns; thin lines represent the non-coding regions. Dotted lines connect the *Cs* regions of high similarity between the two species. The two thick lines in the lower part indicate regions that we have sequenced in *D. melanogaster* and *D. simulans*; the rest of the melanogaster sequence is from Marsh et al. (1986) and Eveleth and Marsh (1987). The gene structure and arrangement are the same in *D. simulans* as in *D. melanogaster*.

ments. Partial sequence of *amd* in *D. simulans* was obtained from yet another clone, which is encompassed by the PCR fragment.

The sequences reported here have been deposited in GenBank database, accession numbers AF091327, AF091328, AF091329, AF121109.

2.3. Alignment and analysis

The sequences were edited and assembled using programs of the Fragment Assembly module of the GCG package (Wisconsin Package Version 9.1). Various GCG programs were also used for alignment and translation. Inference about coding regions was primarily obtained by comparison of the *S. lebanonensis* and *D. melanogaster* sequences seeking regions of high similarity. Additionally, the programs GENIE (Reese et al., 1997) and FGENED (Solovyev et al., 1994) were used for predicting putative exons. Analysis of codon preference was performed with the CODONPREFERENCE program of the GCG package which implements the method of Gribskov et al. (1984). A Fourier transform analysis was performed using the Fast Fourier Transform of the computer program Origin (version 4.10, Microcal Software, Inc.). This method unveils periodicity patterns along binary strings. Such strings were created by using a 1 at each substituted position, and a 0 at identical positions. In addition to the aligned coding regions of *amd*, *Ddc*, and *Cs* of *D. melanogaster*, *D. simulans*, and *S. lebanonesis*, we also used for illustrative purposes *hsr-omega* exons 1 and 2 of *D. melanogaster* (U18307) and *D. pseudoobscura* (X16337). Codon-use bias was assessed by estimating ENC, the 'effective number of codons' (Wright, 1990). Higher values of ENC correspond to lower codon-use bias. Heterogeneity of substitutions along amino acid sequences was tested with the unmodified variance test of Goss and Lewontin (1996). 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, *Scaptodrosophila lebanonensis*, *D. melanogaster*, *D. simulans*, *D. immigrans*, *D. mimica*, *D. (Scaptomyza) palmae*, and *D. (Samoaia) 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,

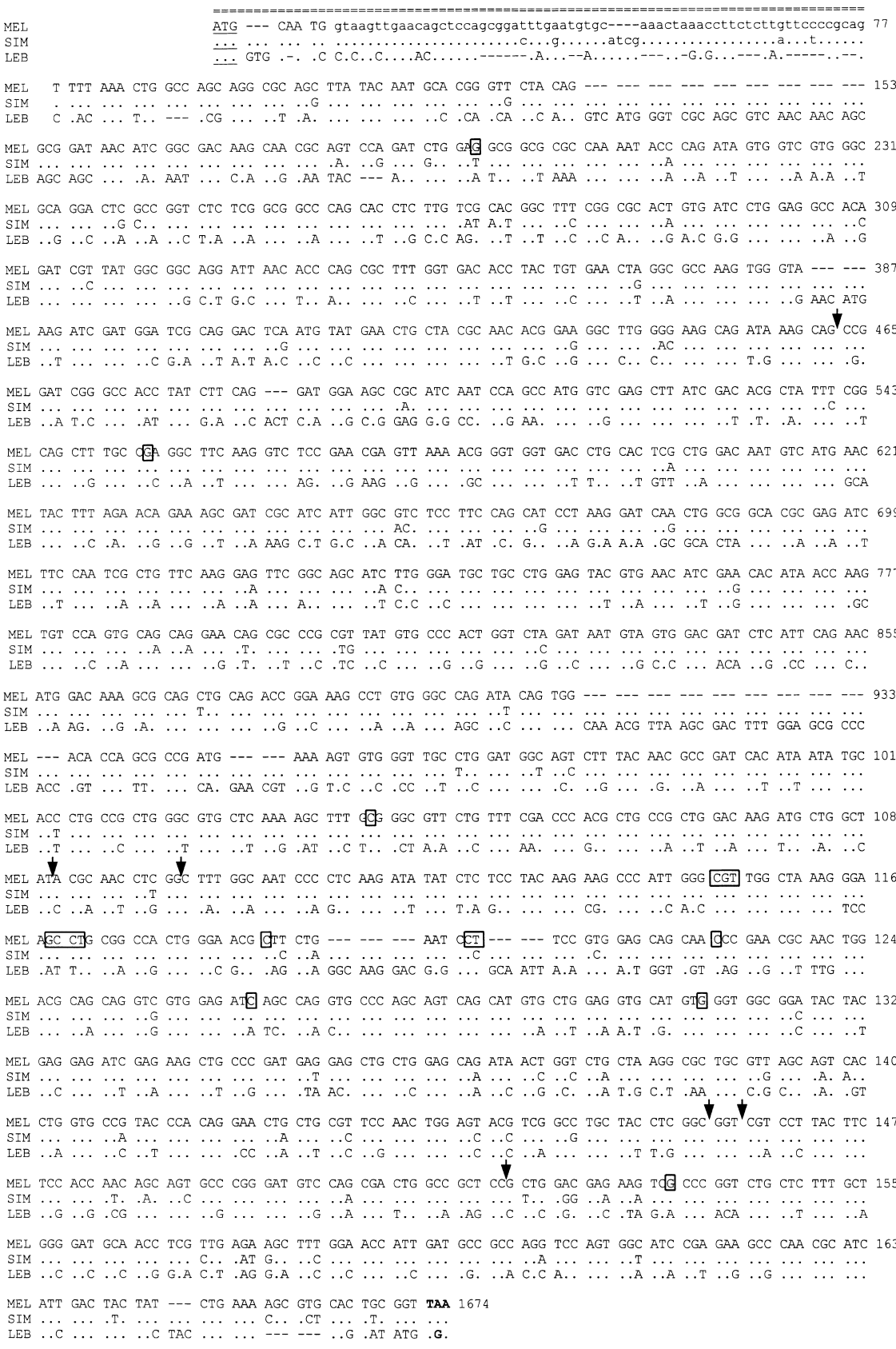

Fig. 2. Alignment of the *Cs* coding region between *D. melanogaster* (MEL), *D. simulans* (SIM), and *S. lebanonensis* (LEB). The intron in *D. melanogaster* and *D. simulans* is shown in lowercase letters at the top of the figure; the proposed initiator ATGs are underlined; the stop codons are in bold. 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.

1987). 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 coding 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 (Gribskov 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

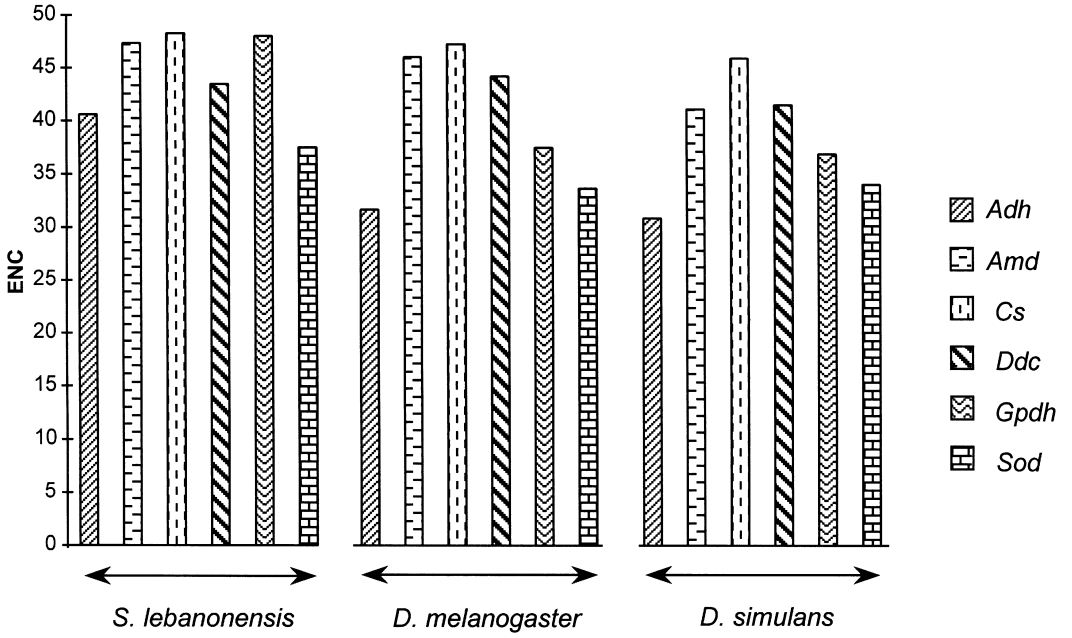

Fig. 3. Codon usage bias in six genes in *S. lebanonensis*, *D. melanogaster*, and *D. simulans*. A larger effective number of codons (ENC) indicates lesser codon usage bias.

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 *S. lebanonensis*, but 504 amino acids in *D. melanogaster* and *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 by Eveleth and Marsh, 1987; but see Okuyama et al., 1997). The correct coding region of *Cs* that we have now determined in *D. melanogaster* makes the gene density even more spectacular, with the stop codons of *Ddc* and *Cs* genes being only 366 bp apart. Our study shows that the tight packing also occurs in *S. lebanonensis*, in which the *amd* stop codon is just 686 bp from the *Cs* start codon, and the stop codons of *Cs* and *Ddc* are only 722 bp apart from one another (Fig. 1). The suggestion that mutagenic silence of the *Cs* may have occurred in *D. melanogaster* as a consequence of evolutionarily recent modifications in the gene's structure (Eveleth and Marsh, 1987) becomes unconvincing, given that *Cs* has remained tightly packed with *Ddc* and *amd*

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)

		melanogaster–simulans	*melanogaster–lebanonensis*	*simulans–lebanonensis*
Amd	n-syn	0.010 ± 0.004	0.105 ± 0.013	0.099 ± 0.012
	syn	0.151 ± 0.027	1.313 ± 0.174	1.372 ± 0.188
	ratio	0.066	0.080	0.072
Cs	n-syn	0.027 ± 0.005	0.212 ± 0.016	0.217 ± 0.016
	syn	0.149 ± 0.022	1.562 ± 0.204	1.467 ± 0.186
	ratio	0.181	0.136	0.148
Ddc	n-syn	0.003 ± 0.002	0.066 ± 0.010	0.064 ± 0.010
	syn	0.064 ± 0.018	1.239 ± 0.175	1.155 ± 0.156
	ratio	0.047	0.053	0.055
Adh	n-syn	0.002 ± 0.002	0.101 ± 0.016	0.103 ± 0.017
	syn	0.052 ± 0.021	0.802 ± 0.122	0.765 ± 0.117
	ratio	0.038	0.126	0.135
Gpdh	n-syn	0.000 ± 0.000	0.012 ± 0.005	0.014 ± 0.005
	syn	0.060 ± 0.019	1.296 ± 0.251	1.194 ± 0.209
	ratio	0.000	0.009	0.012
Sod	n-syn	0.000 ± 0.000	0.113 ± 0.020	0.108 ± 0.019
	syn	0.114 ± 0.037	1.508 ± 0.473	1.938 ± 1.030
	ratio	0.000	0.075	0.056

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

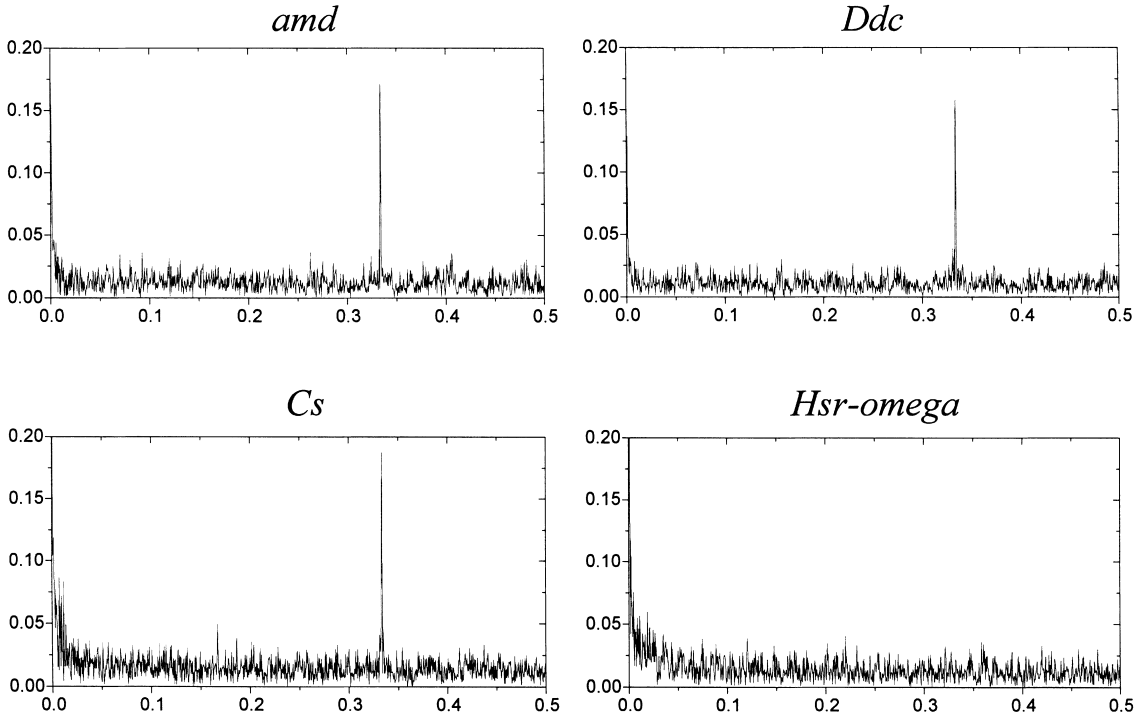

Fig. 4. Fourier transform of the substitution pattern in four genes: *amd*, *Ddc*, and *Cs* are from *D. melanogaster*, *D. simulans*, and *S. lebanonensis*, *hsr-omega* is from *D. melanogaster* and *D. pseudoobscura*. A dominant substitution frequency of 1/3 is revealed for the exon sequences of *amd*, *Ddc*, and *Cs*, while no predominant peak is observed for the two exons of the non-coding *hsr-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 sequencies 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 untypical 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; Stathakis 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)

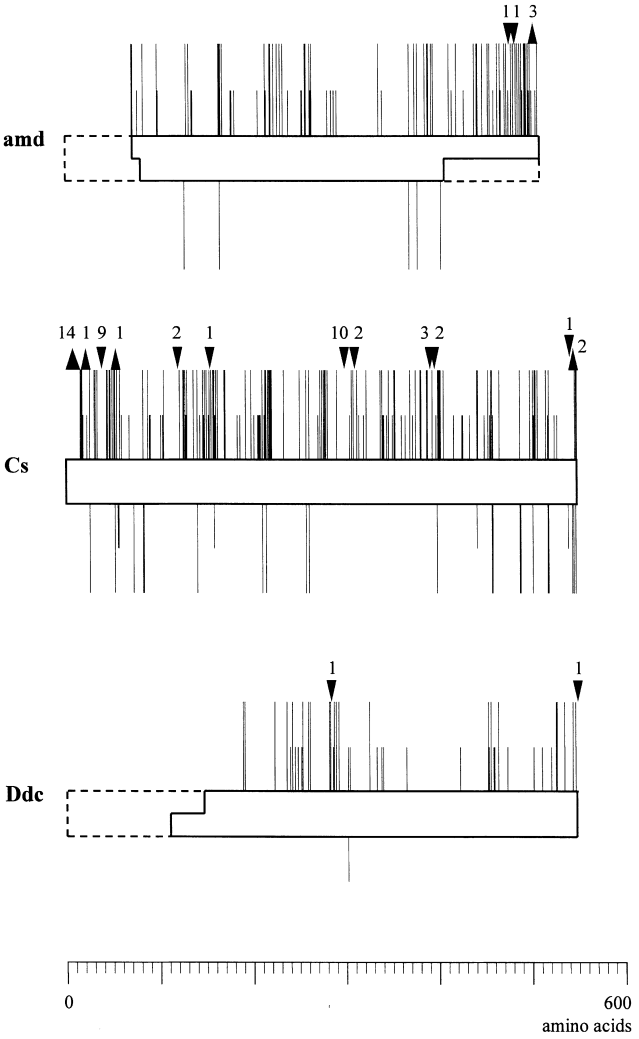

Fig. 5. Amino acid substitutions along *amd*, *Cs*, and *Ddc* between *D. melanogaster*, and *D. simulans* or *S. lebanonensis*. The three boxes represent, from top to bottom respectively, the aligned protein sequences of *amd*, *Cs*, and *Ddc*. The upper part of the box corresponds to the alignment between *D. melanogaster* and *S. lebanonensis*, the lower part between *D. melanogaster* and *D. simulans*. Dashed lines indicate regions where the alignment was not feasible owing to the absence of at least one sequence. Substitutions are shown by vertical lines: short when the amino acid replacements are conservative (D/E, K/R/H, N/Q, S/T, I/L/V, F/W/Y, or A/G, according to Smith and Smith, 1990), and long when they are not conservative (any other substitutions). Arrowheads indicate the position of gaps, pointing down when they occur in *D. melanogaster*, and up for the compared sequence. Numbers indicate the number of amino acid residues. A scale is at the bottom.

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,

compared with random distribution, of large segments in *Ddc* and *amd*, and of contiguous substitutions in *Cs* (data not shown). A remarkable area of low contrained evolution is the carboxyl end of *amd*, while the central areas of *amd* and *Ddc* appear to be the most constrained ones. In *Cs* it is difficult to distinguish areas of low interspecific variation, and the non-random distribution of substitutions is probably due to an excess of runs of contiguous substitutions, as mentioned above.

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). No single sequence that is particularly close to the *Cs* gene could be singled out; instead, all these sequences are approximately equally similar to it, suggesting that the split of *Cs* from the common ancestral gene is very ancient, perhaps predating the diversification of the major multicellular kingdoms.

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 protein 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*. *Ddc* 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. *Cs* is affected the most by the substitution process, with runs of contiguous substitutions along its whole length.

(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

Corticosteroid-binding protein in yeast (*Candida albicans*)	Polyamine oxidase in maize (*Zea mays*)
leban: 55 SAKQNTQIVVIGAGLAGLSAAQHLLRHGFRS---TIVLEATDRYGGRVNS-----KRFGD 106 S ++T++++IGAG++GL AA+ +L F + +V+EA +R GGR+ + + G yeast: 2 SRTKSTKVLIIGAGVSGLKAAETILSKSFLTGDDVLVVEAQNRIGGRLKTTDTSQSKLGI 61	leban: 50 QYNLESAKQNTQIVVIGAGLAGLSAAQHLLRHGFRSTIVLEATDRYGGRVNSKRFGDTYC 109 Q+ +A +++V+GAG++G+SAA+ L G ++LEATD GGR++ F maize: 23 QHGSLAATVGPRVIVVGAGMSGISAAKRLSEAGITDLLILEATDHIGGRMHKTNFAGINV 82
leban: 107 TYCELGAKWVNMNIDGAHNTIYELLRNAEGLRKQLKQ-----RECANYVHTQGREVPPNM 161 Y +LGA W + D +N + + N +GL K ++ + T EVP yeast: 62 NY-DLGASWFH---DSLNNIVLNHMIN-DGLLDDEKDVYFDDKDLKTFSSTG--EVP--- 111	leban: 110 ELGAKWVNMNIDGAHNTIYELLRNAEGLRK------QLKQRECANYVHTQGREVPPNMVE 162 ELGA WV G N I+ ++ + LR L Q + +E maize: 83 ELGANWVEGVNGGKMNPIWPIVNSTLKLRNFRSDFDYLAQNVYKEDGGVYDEDYVQKRIE 142
leban: 162 VELIDMQFRQLCRGFKVSEKVKSGGDLHVLDNVMAYFKTESEKLVGHSYPDPEKRALARE 221 ++D + + VL+++ Y + + +G PD R + + yeast: 112 --IVDKKLNR----------VLEDIEKYIQLYFNRNLG--VPDLSLRDIVAQ 149	leban: 164 LIDMQFRQLCRGFKVSEKVKSGG--DLHVL----------DNVMAYFKTESEK 204 L D G K+S + + G D+ +L D V+ Y+K + E maize: 143 LADSVEEM---GEKLSATLHASGRDDMSILAMQRLNEHQPNGPATPVDMVVDYYKFDYE- 198
leban: 222 IFQS----LFKEFSSILGCCLEYVNI------EHITS--CPVQQELRPLYVPTGLDNVLD 269 F+ + +E G + Y+ + I+ + R L G +++ yeast: 150 YFEKYNRLITEEQREYCGRMMRYLEFWFGISWDRISGKYAVTTHQGRNLLNKKGYGYLVE 209	leban: 205 LVGHSYPDPEKRALAREIFQSLFKEFSSILGCCLEYVNIEHITSCPVQQELRPLYVPTGL 264 P +L + + F +F + + E + +Y G maize: 199 ----FAEPPRVTSLQNTVPLATFSDFGDDVYFVADQRGYEAV----------VYYLAGQ 243
leban: 270 TLTQHISKEQLQTGKPVGSIQWQTLSDFGAPTSPLPQERKCVACLDGTLYSADHIICTLP 329 +L + I + L +PV I + D G +R V ++G D++I T+P yeast: 210 SLAKRIPESSLLLEEPVNKII-RNNKDAG--------KRVLVETINGLQIFCDYLIVTVP 260	leban: 265 DNVLDTLTQHISKEQLQTGKPVGSIQWQTLSDFGAPTSPLPQERKCVACLDGTLYSADHI 324 D + I +LQ K V I++ G T V D ++YSAD++ maize: 244 YLKTDDKSGKIVDPRLQLNKVVREIKYSP----GGVT---------VKTEDNSVYSADYV 290
leban: 330 LGVL----KNFSAILFKPALPLEKLQAIRNLGYGNPVKIYLAYKRPISRWLKSNLRPLGA 385 +L + +I ++P LP +++I ++ +G K+ + R K + + yeast: 261 QSILLLEESSPYSIKWEPKLPQRLVESINSIHFGALGKVIFEFDRIFWDNSKDRFQIIAD 320	leban: 325 ICTLPLGVLKNFSAILFKPALPLEKLQAIRNLGYGNPVKIYLAYKRPISRWLKSNLRPLG 384 + + LGVL++ I FKP LP K++AI KI+L + R W + R maize: 291 MVSASLGVLQS-DLIQFKPKLPTWKVRAIYQFDMAVYTKIFLKFPRKF--WPEGKGR--- 344
leban: 386 QLGKDEPAITVNGRQERLWTQQVVEISQLPSSQHVLEIRVGGGYYDEIEKLPDVTLLEQI 445 D + + V ++ + + L I + +E PD yeast: 321 HTDGDLSRELTELPKPFTYPLFAVNFGRVHNGKASLVILTQAPLTNYLETHPDQAWQYYQ 380	leban: 385 AQLGKDEPAITVNGRQERLWTQQVVEISQLPSSQHVLEIRVGGGYYDEIEKLPDVTLLEQ 444 E + + R+ Q E Q P + +VL + V IE+ D + maize: 345 ------EFFLYASSRRGYYGVWQEFE-KQYPDA-NVLLVTVTDEESRRIEQQSDEQTKAE 396
leban: 446 TALLRQCLRNRLVPYPQALLRSNWSTSACYLGGRPYFSTTSSARDV----QRLAEPLGDI 501 L + + + +P P + ++W+T+ G T D+ E LG + yeast: 381 PMLQKLSINDEPIPDPINTIVTDWTTNPYIRGSYSTMYTNDDPSDLIISLSGDFEDLGIL 440	leban: 445 ITALLRQCLRNRLVPYPQALLRSNWSTSACYLGGRPYFSTTSSARDVQRLAEPLGDIAPT 504 I +LR+ + VP +L W + Y G + + + +L P+G + maize: 397 IMQVLRKMFPGKDVPDATDILVPRWWSDRFYKGTFSNWPVGVNRYEYDQLRAPVGRV--- 453
leban: 502 APTLLFAGDATALKGFGTIDGARTSGIREAQRIID 536 P + FAG+ T +G G + GA SGI A I++ yeast: 441 EPYIKFAGEHTTSEGTGCVHGAYMSGIYAADCILE 475	leban: 505 LLFAGDATALKGFGTIDGARTSGIREAQRIID 536 F G+ T+ G + GA SGI A+ +I+ maize: 454 -YFTGEHTSEHYNGYVHGAYLSGIDSAEILIN 484

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.

codon usage bias, suggesting that *Cs* is not functionally so highly constrained as the other genes.

(5) The *Cs* 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

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

Cooley, J.W., Tukey, J.W., 1965. An algorithm for the machine calculation of complex Fourier series. Math. Comput. 19, 297–301.

Eveleth, D.D., Marsh, J.L., 1986. Evidence for evolutionary duplication of genes in the dopa decarboxylase region of *Drosophila*. Genetics 114, 469–483.

Eveleth, D.D., Marsh, J.L., 1987. Overlapping transcription units in *Drosophila*: sequence and structure of the *Cs* gene. Mol. Gen. Genet. 209, 290–298.

Eveleth, D.D., Gietz, R.D., Spencer, C.A., Nargang, F.E., Hodgetts, R.B., Marsh, J.L., 1986. Sequence and structure of the dopa decarboxylase gene of *Drosophila*: evidence for novel RNA splicing variants. EMBO J. 5, 2663–2672.

Fini, M.E., Bendena, W.G., Pardue, M.L., 1989. Unusual behavior of the cytoplasmic transcript of *hsr omega*: an abundant, stress-inducible RNA that is translated but yields no detectable protein product. J. Cell Biol. 108, 2045–2057.

Goss, P.J.E., Lewontin, R.C., 1996. Detecting heterogeneity of substitution along DNA and protein sequences. Genetics 143, 589–602.

Gribskov, M., Devereux, J., Burgess, R.R., 1984. The codon usage plot: graphic analysis of protein coding sequences and prediction of gene expression. Nucleic Acids Res. 12, 539–549.

Jackson, F.R., 1990. Prokaryotic and Eukaryotic pyridoxal-dependent decarboxylases are homologous. J. Mol. Evol. 31, 325–329.

Kwiatowski, J., Skarecky, D., Bailey, K., Ayala, F.J., 1994. Phylogeny of *Drosophila* and related genera inferred from the nucleotide sequence of the Cu,Zn *Sod* gene. J. Mol. Evol. 38, 443–454.

Kwiatowski, J., Krawczyk, M., Jaworski, M., Skarecky, D., Ayala, F.J., 1997. Erratic evolution of glycerol-3-phosphate dehydrogenase in *Drosophila*, *Scaptomyza*, and *Ceratitis*. J. Mol. Evol. 44, 9–22.

Li, W.-H., 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. 36, 96–99.

Li, W.-H., 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA.

Maroni, G., 1993. An Atlas of *Drosophila* Genes: Sequences and Molecular Features. Oxford University Press, New York.

Marsh, J.L., Erfle, M.P., Leeds, C.A., 1986. Molecular localization, developmental expression and nucleotide sequence of the *alpha-methyldopa hypersensitive* gene of *Drosophila*. Genetics 114, 453–467.

Moriyama, E.N., Hartl, D.L., 1993. Codon usage and base composition of nuclear genes in *Drosophila*. Genetics 134, 847–858.

Moriyama, E.N., Powell, J.R., 1997. Codon usage bias and tRNA abundance in *Drosophila*. J. Mol. Evol. 45, 514–523.

Okuyama, E., Tachida, H., Yamazaki, T., 1997. Molecular analysis of the intergenic region of the duplicated *Amy* genes of *Drosophila melanogaster* and *Drosophila teissieri*. J. Mol. Evol. 45, 35–42.

Palumbi, S., Martin, A., Romano, S., MacMillan, W.O., Stice, L., Grabovsky, G., 1991. The Simple Fool's Guide to PCR, Version 2.0. Department of Zoology and Kewalo Marine Laboratory, University of Hawaii, Honolulu.

Reese, M.G., Eeckman, F.H., Kulp, D., Haussler, D., 1997. Improved splice site detection in Genie. J. Comput. Biol. 4, 311–323.

Russo, C.A.M., Takezaki, N., Nei, M., 1995. Molecular phylogeny and divergence times of drosophilid species. Mol. Biol. Evol. 12, 391–404.

Shields, D.C., Sharp, P.M., Higgins, D.G., Wright, F., 1988. 'Silent' sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. 5, 704–716.

Smith, R.F., Smith, T.F., 1990. Automatic generation of primary sequence patterns from sets of related protein sequences. Proc. Natl. Acad. Sci. USA 87, 118–122.

Solovyev, V.V., Salamov, A.A., Lawrence, C.B., 1994. Predicting internal exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. Nucleic Acids Res. 22, 5156–5163.

Spencer, C.A., Gietz, R.D., Hodgetts, R.B., 1986a. Overlapping transcription units in the *Dopa decarboxylase* region of *Drosophila*. Nature 322, 279–281.

Spencer, C.A., Gietz, R.D., Hodgetts, R.B., 1986b. Analysis of the transcription unit adjacent to the 3′ end of the dopa decarboxylase gene in *Drosophila melanogaster*. Dev. Biol. 114, 260–264.

Stathakis, D.G., Pentz, E.S., Freeman, M.E., Kullman, J., Hankins, G.R., Pearlson, N.J., Wright, T.R.F., 1995. The genetic and molecular organization of the *Dopa decarboxylase* gene cluster of *Drosophila melanogaster*. Genetics 141, 629–655.

Tatarenkov, A., Kwiatowski, J., Skarecky, D., Barrio, E., Ayala, F.J., 1999. On the evolution of *Dopa decarboxylase* (*Ddc*) and *Drosophila* systematics. J. Mol. Evol.. in press

Tsonis, A.A., Elsner, J.B., Tsonis, P.A., 1991. Periodicity in DNA coding sequences: implications in gene evolution. J. Theor. Biol. 151, 232–331.

Wang, D., Marsh, J.L., Ayala, F.J., 1996. Evolutionary changes in the expression pattern of a developmentally essential gene in the three *Drosophila* species. Proc. Natl. Acad. Sci. USA 93, 7103–7107.

Wisconsin Package Version 9.1, Genetics Computer Group (GCG), Madison, WI.

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

Wright, T.R.F., 1996. Phenotypic analysis of the *Dopa decarboxylase* gene cluster mutants in *Drosophila melanogaster*. J. Hered. 87, 175–190.