Analysis of nucleotide substitutions and amino acid conservation in the *Drosophila Adh* genomic region

R. Albalat, G. Marfany & R. González-Duarte*
Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Spain *Author for correspondence

Received 14 February 1994 Accepted 3 June 1994

Key words: Adh, Adh-dup, Drosophila, molecular clock, nucleotide substitution rates

Abstract

The homologous genomic region that contains two paralogous genes, *Adh* and *Adh-dup*, was compared in several *Drosophila* species. Sequences were analyzed as follows: a) At the nucleotide level, *Ka* and *Ks* values were determined for each pair of species. *Ka-Adh* and *Ka-Adh-dup* are not significantly different. However, *Ks-Adh* values are significantly lower than *Ks-Adh-dup*, which are more variable. In agreement with other reports, lower *Ks* values for *Adh* correlate with a high level of gene expression and relatively high percentage of G+C content in the third codon position, while the opposite applies to *Adh-dup*. b) At the protein level, amino acid comparisons reveal conserved regions shared by ADH and ADH-DUP, which have been assigned to known functional domains. Key residues for dehydrogenasic function are also found in ADH-DUP, thus pointing to a dehydrogenase activity for ADH-DUP, albeit very different from that of ADH.

Introduction

Evolutionary processes leave their imprint on the genetic material of living organisms. For many years, the long-distance effects of such changes could be evaluated only at the morphological and cytological level. Now, molecular data contribute to our understanding of the forces underlying the dynamics of species and therefore the perspective of the analysis of long-term evolution of genes and genomes has changed considerably.

The alcohol dehydrogenase (*Adh*) gene has been thoroughly characterized in many *Drosophila* species. Moreover, an *Adh-dup* sequence was reported downstream from *Adh*, almost overlapping its 3’ terminal region. These genes are suitable for comparative evolutionary studies as they share a common ancestor and retain high sequence similarity (Schaeffer & Aquadro, 1987).

Analysis of synonymous and non-synonymous nucleotide substitutions (*Ks* and *Ka* respectively) in a wide variety of protein coding sequences has shown that each gene evolves at a specific rate (Li & Graur, 1991). For all the genes analyzed, the synonymous substitution rate greatly exceeds the non-synonymous rate. Because the relative proportion of *Ks* and *Ka* varies among genes, one way to approach the evolution of a particular genomic region is to compute this ratio, as has already been done with *Adh* in some *Drosophila* species (McDonald & Kreitman, 1991; Ohta, 1993) as well as in several mammalian genes (Whitfield, Lovell-Badge & Goodfellow, 1993; Tucker & Lundrigan, 1993). According to the molecular clock hypothesis, *Ka* and *Ks* values for any gene in any pair of species should correlate with evolutionary time. Nevertheless, some genes are reported to evolve differently, with a *Ka/Ks* ratio that is much higher than expected in comparison with other genes of the same species (Hill & Hastie, 1987; Hughes & Nei, 1988; Hughes, Ota & Nei, 1990; Whitfield, Lovell-Badge & Goodfellow, 1993; Tucker & Lundrigan, 1993).

We have calculated and analyzed *Ks* and *Ka* for *Adh* and *Adh-dup* in twelve *Drosophila* species using the LWL91 program (Li, Wu & Lou, 1985). Our results show that non-synonymous substitutions are consistent with the elapsed time, while the rate of synonymous substitutions may vary in a species-dependent manner and correlate to the degree of codon usage bias and the
level of expression of each gene, in agreement with other reports (Shields et al., 1988; Sharp & Li, 1989; Moriyama & Gojobori, 1992).

On the other hand, the comparison of ADH and ADH-DUP at the amino acid level facilitates the definition of some conserved structural domains and shows that, in spite of their different evolutionary pattern, residues that are essential for dehydrogenase function have been maintained in the ADH-DUP protein.

Material and methods

All the DNA sequences used in this study were obtained from GenEmbl database. The Adh and Adh-dup genes of D. ambigua, D. guanche, D. immigrans, D. lebanonensis, D. madeirensis and D. subobscura had been sequenced in our laboratory. The accession numbers are: D. ambigua X54813, D. guanche X60112, D. immigrans M97638, D. lebanonensis M97637, D. madeirensis X60113, D. mauritiana M19264, D. melanogaster M17827, D. miranda M60998, D. persimilis M60997, D. pseudoobscura Y00602, D. subobscura M55545 and D. teissieri X54228.

The coding regions of Adh and Adh-dup were used to calculate the Ks and Ka values (synonymous and non-synonymous nucleotide substitutions respectively) using the LWL91 program (Li, Wu & Lou, 1985). In Adh-dup comparisons, the shorter sequence of the pair was always taken as reference for nucleotide alignments. To compare the Ka/Ks ratios of the two genes among groups we used the non-parametric Mann-Whitney test, which is based on the assumption that the two sets of data analyzed have distributions of the same shape. These two sets of samples (Ka/Ks ratios for Adh and Adh-dup) are considered independent as the evolutionary pattern may differ between genes and does not depend on phylogenetic distances. Besides, assuming that the evolutionary pattern of a specific gene was maintained through time, the value produced would be a constant, irrespective of the species compared.

However, to compare either Ka or Ks values between Adh and Adh-dup, we used the hierarchical sampling scheme represented in Fig. 1, where we consider species from three different subgenera: Scaptodrosophila, Drosophila and Sophophora. All of them split up approximately at the same time, very close to the origin of the genus Drosophila. We considered one of the subgenera, Scaptodrosophila, and its representative, D. lebanonensis, as the reference species, since the phylogenetic distance to the other species is nearly equivalent. This data set for the parameters Ka and Ks (with D. lebanonensis as the outgroup) is an unbalanced 2-way nested classification, adjusted to the model:

\[ y_{ijl} = \mu + g_i + s_j(i) + e_{ijl} \]

where \( \mu \) is the overall grand mean, \( g_i \) is the fixed effect of the ith gene (i = 1, 2), \( s_j(i) \) is the random effect of the jth subgroup within the gene i and \( e_{ijl} \) is the residual error associated with the Ka(Ks) value of the lth species within the jth subgroup. Searle, Casella and McCulloch (1992) suggested maximum likelihood (ML) or restricted maximum likelihood to analyze unbalanced data. In this case, \( F \) ratios for average comparisons of Ka(Ks) between Adh and Adh-dup are based on the asymptotic variance-covariance matrix (REML method). The program 3V of the BMPD Statistical Software package (1992, University of California, Berkeley, California), implemented on a VAX-6610 VMS at the Centre de Calcul de la Universitat Autonoma de Barcelona, was used for data analysis. One of the requirements of ML or REML is the assumption of an underlying probability distribution (normal distribution in this case), although their properties seem to be sufficiently reliable.

A statistical analysis of the differences in the G+C percentage in the third codon position (excluding Trp and Met) between Adh and Adh-dup was performed with the REML method, but using the transformation of percentages into the arcsine of the square root of each value.

The primary structures of ADH and ADH-DUP were deduced from the translation of the corresponding coding regions using the ASSEMBLE and TRANSLATE applications of the GCG Software Package (Devereux, Haeberli & Smithies, 1984). The proteins were aligned using the PILEUP application. A two-residue gap corresponding to the second and third amino acids was introduced in order to compensate for the insertion present in the Adh of the melanogaster group.

Results and discussion

Analysis of nucleotide substitution rates: comparison of Adh and Adh-dup sequences

The nucleotide substitutions in Adh and Adh-dup were quantified and the Ka/Ks ratio for both genes in each