Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 9047–9051, September 1995
Evolution

Maintenance of pre-mRNA secondary structure by epistatic selection

(linkage disequilibrium/compensatory mutations)

DAVID A. KIRBY*, SPENCER V. MUSE†, AND WOLFGANG STEPHAN*‡

*Department of Zoology, University of Maryland, College Park, MD 20742; and †Department of Biology and Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park, PA 16802

Communicated by Michael T. Clegg, University of California, Riverside, CA, June 16, 1995

ABSTRACT **Linkage disequilibrium between polymorphisms in a natural population may result from various evolutionary forces, including random genetic drift due to sampling of gametes during reproduction, restricted migration between subpopulations in a subdivided population, or epistatic selection. In this report, we present evidence that the majority of significant linkage disequilibria observed in introns of the alcohol dehydrogenase locus (*Adh*) of *Drosophila pseudoobscura* are due to epistatic selection maintaining secondary structure of precursor mRNA (pre-mRNA). Based on phylogenetic-comparative analysis and a likelihood approach, we propose secondary structure models of *Adh* pre-mRNA for the regions of the adult intron and intron 2 where clustering of linkage disequilibria has been observed. Furthermore, we applied the likelihood ratio test to the phylogenetically predicted secondary structure in intron 1. In contrast to the other two structures, polymorphisms associated with the more conserved stem–loop structure of intron 1 are in low frequency, and linkage disequilibria have not been observed. These findings are qualitatively consistent with a model of compensatory fitness interactions. This model assumes that mutations disrupting pairing in a secondary structural element are individually deleterious if they destabilize a functionally important structure; a second "compensatory" mutation, however, may restabilize the structure and restore fitness.**

The analysis of epistatic interactions has played an important role in population genetics theory since it was introduced by Haldane (1) and Wright (2). Historically, epistatic interactions are defined as interactions between genes. Epistatic interactions are expected to lead to nonrandom associations between polymorphisms at different loci within populations; however, nonrandom associations are rarely detected in natural populations. Most notably, extensive studies of linkage disequilibrium based on allozyme variation at many loci in natural populations of *Drosophila* have failed to lend support to Wright's ideas (3, 4). The prevailing view is that the lack of significant associations could be the result of the large map distances between most of the loci surveyed and/or the low density of selection per map unit (4–6). An alternative view is that the power of the statistical tests used to detect linkage disequilibrium between allozyme loci was too low and that moderate levels of disequilibrium between rather loosely linked allozyme loci can be detected by using a more powerful statistical approach (7).

In contrast, the observations emerging from the application of recombinant DNA technology to *Drosophila* population genetics show several examples of extensive nonrandom associations between DNA polymorphisms over relatively short distances. In restriction map surveys of natural populations of *Drosophila melanogaster* and *Drosophila simulans*, nonrandom associations between pairs of polymorphisms have been found over tens of kilobases (8, 9). Linkage disequilibria tend to decay as the distance between the compared sites increases. However, recombination typically does not break up all pairs of polymorphisms, thus creating a scattering of disequilibria. This seemingly random distribution of linkage disequilibria makes a molecular analysis difficult. There are, however, regions in which the pattern of genetic correlations is more regular, e.g., the white locus of *D. melanogaster* (10, 11) and the *Adh* locus in *Drosophila pseudoobscura* (12). In these two gene regions, strong linkage disequilibria were clustered within the transcriptional unit.

Standard statistical tests applied to the locus as a whole have failed to suggest the action of past positive Darwinian selection in the *Adh* gene region in *D. pseudoobscura* populations (13). However, two small segments were identified at the *Adh* locus that show strong linkage disequilibrium within each region (12): nt 331–355 of the adult intron and nt 1454–1500 of intron 2. Together, these two clusters contain almost 90% of the statistically significant disequilibria found within the entire *Adh* locus. Since within each cluster almost all disequilibria show a consistent pattern between subpopulations with regard to strength and direction of association, Schaeffer and Miller (12) concluded that these correlations are due to epistatic selection (14, 15) rather than random genetic drift and restricted migration (16–18). In this report, we examine the hypothesis that the linkage disequilibria in the two clusters at the *Adh* locus are caused by epistatic selection maintaining the secondary structures of pre-mRNA in these regions. This work is a continuation of our efforts (19) to infer secondary structure models of the more conserved portions (e.g., exons) of the *Drosophila Adh* locus.

In it simplest form, the mechanism underlying the action of epistatic selection on secondary structure may be as follows: A mutation occurring in a secondary structural element such as the helix of an RNA hairpin may be individually deleterious because it increases the structure's free energy, which may destabilize this structure. However, the pairing potential of a functionally important structure, and thus fitness, can be restored if a second "compensatory" mutation occurs in the complementary sequence of the helix. To examine this hypothesis, we first inferred pre-mRNA secondary structures in the two regions of interest based on phylogenetic comparisons (20, 21). Then, we tested the significance of the phylogenetically predicted stems by using the likelihood approach of Muse (22). To complete our analysis of intron pre-mRNA structures, we reanalyzed the hairpin structure in intron 1 we have inferred (19) from phylogenetic comparisons by subjecting it to the likelihood ratio test (LRT).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: LRT, likelihood ratio test; ha1 and ha2, haplotype 1 and haplotype 2, respectively.

‡To whom reprint requests should be addressed.

MATERIALS AND METHODS

Phylogenetic Comparison. The inference of an RNA secondary structure from DNA sequence comparison is usually based on the Woese–Noller criterion (20, 21); i.e., a putative helix of a RNA structural element is considered "proven" if two or more covariations, caused by independently occurring base substitutions in the complementary sequences of a putative helix, are detected in sequence comparisons. However, this is only a heuristic criterion that does not account for divergence levels and the number of species in the comparisons (19, 22). Therefore, it must be complemented by more rigorous statistical tests (see below) or experimental procedures. To meet the Woese–Noller criterion, we have done sequence comparisons between the following species with various genetic distances to *D. pseudoobscura; Drosophila persimilis, Drosophila miranda, Drosophila ambigua, D. melanogaster, Drosophila teissieri, Drosophila erecta*, and *Drosophila lebanonensis.* The first two are sibling species of *D. pseudoobscura*; the more distantly related *D. ambigua* is also a member of the *obscura* group. *D. melanogaster, D. teissieri*, and *D. erecta* are from the *melanogaster* species group. Both groups belong to the subgenus *Sophophora. D. lebanonensis* is from the subgenus *Scaptodrosophila.* Species from the subgenus *Drosophila* could not be included in the comparison of the adult intron and intron 2 sequences (for reasons, see ref. 19 and below).

An integral part of the phylogenetic method is the alignment of the DNA sequences. We used the progressive alignment procedure proposed by James *et al.* (23). This method ties the alignment and inference steps together. It is particularly useful for the phylogenetic comparison of DNA sequences that are diverged, as is the case for the intron sequences. The alignment of homologous nucleotides in the seven sequences of the adult intron is shown in Fig. 1*A*. First, sequences in this region were aligned within the *obscura* and *melanogaster* groups separately. These partial alignments were unambiguous, except for a short repetitive motif in *D. ambigua* (see Fig. 1*A*). Then, sequence alignments were constructed between these species groups and *D. lebanonensis.* In this intergroup alignment, emphasis was placed on features that are conserved within a group. Alignment began with the conserved endpoints of these homologous sequences and then proceeded toward the middle. Once the conserved pairs were aligned, identification of putative pairing regions was used to refine the intergroup alignment. In other words, the putative pairings were aligned between groups to resolve the intergroup alignment. This procedure has also been used in our previous study (19) to align the sequences in intron 1.

In intron 2, the progressive alignment method was applied with some modifications. Intron 2 shows too much divergence to allow the simultaneous alignment of all seven DNA sequences. Only the sequences of the three species of the *obscura* group could be aligned over the entire length of intron 2 (Fig. 2*A*). This was done by aligning first the *persimilis* and *miranda* sequences (in one group) and by resolving the intergroup alignment between this group and *ambigua* with information about putative pairings. In addition, the *D. melanogaster* sequence could be aligned in the small segment encompassing the branchpoint sequence CTAA (position 1483; coordinates from *D. pseudoobscura*) and around the 5′ and 3′ splice sites. Although the branchpoint sequence is an important splicing signal, it is not strongly conserved in most eukaryotes (the *Drosophila* consensus sequence is CTAA). In *D. melanogaster* and in the other three species in Fig. 2*A*, the branchpoint could be unambiguously identified, by using the methods of Mount *et al.* (26). Once the alignments were defined, complementary regions within each sequence were examined for covariations.

Statistical Test of Secondary Structure. The approach of Muse (22) was used to test for the existence of secondary structures predicted by the phylogenetic method. Muse (22) presented an evolutionary model that incorporated the effects of secondary structure in terms of a pairing parameter λ. Pairs of nucleotides are the evolutionary unit, and the 16 × 16 instantaneous substitution rate matrix (**R**) is described in its simplest form as follows:

$$R_{ij} = \begin{cases} \frac{1}{4}\mu\lambda, \text{ 1 difference, pairing gained (e.g., } i = \text{AC}, j = \text{AT}) \\ \frac{1}{4}\mu, \text{ 1 difference, pairing unchanged (e.g., } i = \text{AC}, j = \text{AG}) \\ \frac{1}{4}\mu/\lambda, \text{ 1 difference, pairing lost (e.g., } i = \text{AT}, j = \text{AC}) \\ 0, \text{ 2 differences (e.g., } i = \text{AC, j = TG}) \end{cases} \quad [1]$$

By design, this model reduces to the independent sites, Jukes–Cantor model with substitution rate μ when $\lambda = 1$. In our analysis, the extended Jukes–Cantor model (27) was used,

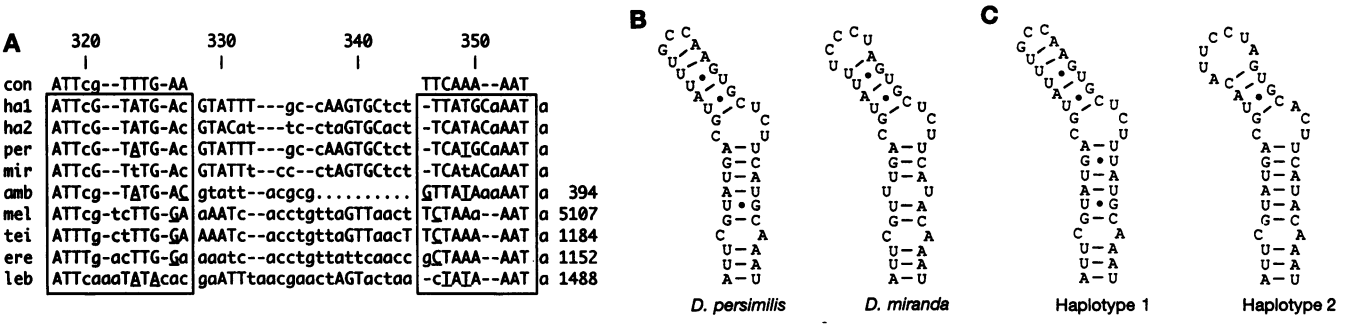

FIG. 1. Sequence alignment and secondary structures for the adult intron. (*A*) Sequence alignment within the adult intron for four species of the *obscura* group (ha1, haplotype 1 of *D. pseudoobscura*; ha2, haplotype 2 of *D. pseudoobscura*; per, *D. persimilis*; mir, *D. miranda*; amb, *D. ambigua*), three species of the *melanogaster* group (mel, *D. melanogaster*; tei, *D. teissieri*; ere, *D. erecta*) and *D. lebanonensis* (leb). Due to a repetitive element within *D. ambigua* (ACGCG), this sequence could only be partially aligned. Coordinates above the alignment are from *D. pseudoobscura* (12), the coordinates after the alignment correspond to the last nucleotide in the alignment for each respective species [GenBank accession nos. M14802 (mel), X54118 (tei), and X54116 (ere); refs. 24 and 25]. Dashes indicate insertion or deletion events and dots indicate portions of sequences that could not be unambiguously aligned. Phylogenetically inferred pairing regions are boxed, and consensus sequences (con) for these regions are shown above the alignment. Consensus sequences are composed of nucleotides that are most commonly paired in the phylogenetically inferred helices. Covariations are underlined and represent deviations away from the consensus sequence. Nucleotides that are involved in pairing regions are capitalized, and nucleotides that remain unpaired are in lowercase type. (*B*) Stem–loop structures for two of the species from the *obscura* group. The drawings were constructed by using LOOPDLOOP (available from D. G. Gilbert via anonymous ftp to ftp.bio.indiana.edu). (*C*) Stem–loop structures for the two haplotypes from *D. pseudoobscura* populations. Note that in both haplotypes there are several differences in the upper stem.

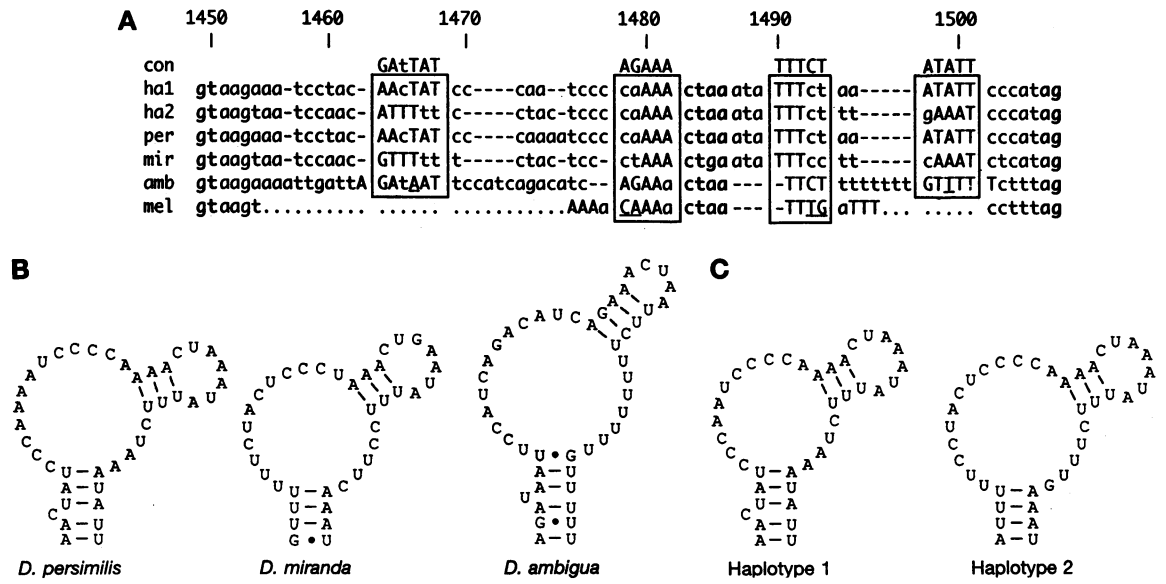

FIG. 2. Sequence alignment and secondary structures for intron 2. (*A*) Sequence alignment of intron 2 for four species from the *obscura* group (ha1, haplotype 1 of *D. pseudoobscura*; ha2, haplotype 2 of *D. pseudoobscura*; per, *D. persimilis*; mir, *D. miranda*; amb, *D. ambigua*) and *D. melanogaster* (mel). The latter sequence could only be partially aligned. The coordinates are from *D. pseudoobscura* (12). Conserved splicing elements are indicated in boldface type (5′ splice site at nt 1449, branchpoint sequence at nt 1483, and 3′ splice site at nt 1507). (*B*) Stem–loop structures of intron 2 for the three *obscura* species showing phylogenetically inferred pairings. (*C*) Stem–loop structures for the two basic haplotypes from *D. pseudoobscura* populations. Haplotype 1 forms a structure similar to that of *D. persimilis*. Haplotype 2 forms a structure similar to that of *D. miranda*. Note that these two structures are significantly different in the lower-pairing region. The branchpoint sequence is in the loop portion of the structures and hence accessible by U2 small nuclear ribonucleoproteins during splicing (for discussion see ref. 19).

which accounts for unequal base frequencies, where λ was defined as in Eq. **1.**

The fact that this model reduces to an independent sites model when $\lambda = 1$ allows the construction of an LRT of the null hypothesis H_O:$\lambda = 1$ (sites evolve independently) vs. the alternative hypothes is H_A:$\lambda > 1$ (pairing is favored). The asymptotic distribution of this LRT statistic is χ^2 with one degree of freedom, and Muse (22) demonstrated that this approximation was good even for potential stem structures of 10 nt. However, the stems we tested are shorter than 10 nt, and the divergence levels are different than those used in Muse (22); so those results may not apply. Additionally, it is difficult to interpret rigorously the *P* values obtained from this test procedure: The locations to be tested were selected by previous analysis of the data to find regions with high levels of complementarity. Thus, we are essentially testing all possible locations for stem structures. To alleviate these difficulties, we used a numerical resampling approach to generate an appropriate null distribution for the LRT statistic. The procedure accounts for sequence length, phylogeny, and observed levels of sequence divergence. Furthermore, the multiple-testing problems inherent with tests of secondary structure are alleviated. The steps of the procedure areas follows (the alignment of the observed sequences is used in all steps): (*i*) With observed data, find the maximum likelihood estimate of λ and the value of LRT as described in ref. 22. (*ii*) Shuffle the observed columns of the alignment. This creates a new set of sequences with the same base frequencies as the observed data but destroys the spatial ordering that provides the secondary structure. (*iii*) Find the consensus sequence for the permuted data. (*iv*) With the MFOLD program of the GCG package (version 7) (28, 29), find the thermodynamically optimal secondary structure of the consensus sequence. (*v*) With the structure from step *iv*, compute and store λ and LRT for the permuted data. (*vi*) Repeat steps *ii–v* 50times. Count the number of permuted data sets that have values of the LRT statistic larger than that of the observed data. For any datasets that have higher LRTs, see if they satisfy the Woese–Noller criterion. (Both of these counts are reported in *Results*.)

RESULTS

In this section we describe the secondary structures detected in the three introns of *Adh* pre-mRNA.

Adult Intron. The clustering of linkage disequilibria occurs in the region of nt 331–355 (12). Our search for a secondary structure concentrated, therefore, on a conserved DNA segment that encompasses this region and is as large as possible. Fig 1*A* shows an alignment of the DNA sequences from seven species (other than *D. pseudoobscura*). These seven sequences were used in the analysis. The phylogenetically supported pairing region is boxed. Four covariations were detected in this set of sequences. This pairing region is conserved in all seven species. It forms the lower part of a stem–loop structure shown in Fig. 1*B* for *D. persimilis* and *D. miranda*. A second helix (nt 328–333/nt 337–342) that forms the upper part of this structure is indicated in Fig. 1*B*. This upper stem may exist but was not supported phylogenetically. No covariations involving Watson–Crick pairs were detected. Only a U·G → A·U change at positions 329/341 was found.

Table 1 shows the results of the LRT. The pairing parameter was estimated as $\lambda = 3.38$, and the LRT statistic was 29.12. None of the 50 simulation runs resulted in higher LRT values for the permuted sequences (the maximum was 17.27). This strongly supports the phylogenetically predicted pairing region of the adult intron.

Two distinctly different (consensus) haplotypes [haplotypes 1 and 2 (ha1 and ha2, respectively)] are segregating in *D. pseudoobscura* populations in nt 331–355. These two haplotypes are aligned in Fig. 1*A*, together with the sequences of the seven species used in the analysis. The frequency of ha1 is 72/99 and the frequency of ha2 is 27/99. Our secondary structure analysis suggests that ha1 and ha2 form helices in the boxed region (see Fig. 1*A*) similar to *D. persimilis*. The helices of these three sequences are identical, except for two pairings: ha1 contains two G·U wobble pairs; in *D. persimilis*, one of the wobble pairs is replaced by a G·C pair, and in ha2, both wobble pairs are replaced by Watson–Crick pairs. Fig. 1*C* shows the secondary structures (lower helices) of the *D. pseudoobscura*

Table 1. Results of LRT procedure

Helix	$\hat{\lambda}$	LRT
Adult intron (lower)	3.38	29.12
Intron 1	4.01	25.21
Intron 2 (upper)	3.94	14.20

LRT procedure (22) was applied to the phylogenetically predicted stem regions and the sequence alignments of Figs. 1*A*, 2*A* (with some modification; see text), and 3. Note that when sites evolve independently, the pairing parameter is $\lambda = 1$; $\lambda > 1$ indicates Watson–Crick base pairing.

haplotypes ha1 and ha2. In addition, Fig. 1*C* shows an upper stem in the ha1 and ha2 structures. As mentioned, above, these upper stems may exist, but they are not supported by our phylogenetic analysis. A total of 12 polymorphisms have been observed between nt 331 and 355 (12). Seven of those form pairs of significant linkage disequilibrium thought to be due to epistatic selection (12). All seven polymorphisms map to the 3′ part (nt 346–350) of the phylogenetically predicted lower stem (2 of 7) and to the upper portion (5 of 7) of the secondary structure (including the bulge loop).

Intron 2. Fig. 2*A* shows an alignment of the entire intron 2 of *D. persimilis, D. miranda*, and *D. ambigua*. The sequence of *D. melanogaster* could be aligned only around the conserved splicing signals; i.e., the 5′ and 3′ splice sites and the branchpoint sequence CTAA (position 1483). A helix (nt 1480–1482/nt 1490–1492) could be inferred in the region encompassing the branchpoint sequence. This pairing region is supported by two covariations among the four sequences compared (Fig. 2*A*). For the statistical test of secondary structure, we used only the portion of intron 2 that could be aligned among these four species. The LRT produced $\lambda = 3.94$ and LRT statistic = 14.20 (Table 1). Simulations on this portion of intron 2 gave the following results: 8 of 50 permutations led to higher test statistics than the observed value, 14.20. However, only 1 of these 8 thermodynamically best potential structures was not eliminated by the Woese–Noller criterion of two covariations. This provides reasonably good support for this pairing region (upper stems in Fig. 2*B*), given that this stem is very short.

We extended our covariation search to the rest of intron 2 (i.e., the part that could only be aligned among the three *obscura* group species). Our phylogenetic analysis seems to suggest a pairing between coordinates 1463–1468 and 1497–1501 (Fig. 2*A*): one covariation was found between *D. persimilis* and *D. ambigua*. In addition, there appears to be a compensatory structural change between *D. persimilis*/*D. ambigua* and *D. miranda*: The sequences of *D. miranda* and *D. persimilis* differ in the boxed region at the 5′ end by a group of three adjacent substitutions and in the boxed region at the 3′ end by two adjacent substitutions; similarly, the sequences between *D. miranda* and *D. ambigua* differ in the 5′ box by two adjacent nucleotides and in the 3′ box even by 4 nt. As a result, homologous nucleotides at a particular stem site seem to pair with different nucleotides in different species. Therefore, our LRT cannot be applied in this case.

Next we consider the haplotypes occurring in natural *D. pseudoobscura* populations. The ha1 and ha2 sequences are aligned with those of the other species in Fig. 2*A*. In the pairing regions, ha1 (frequency: 9/99) is identical with the *persimilis* sequence, and ha2 (frequency: 54/99) is identical with that of *miranda* (except for a G → A replacement at position 1463). Therefore, our phylogenetic analysis suggests that ha1 and ha2 of *D. pseudoobscura* can form secondary structures that are similar to those of pre-mRNAs of *persimilis* and *miranda*. The G (in *miranda*) → A (in *pseudoobscura*) replacement has a stabilizing effect on the structure of ha2 because a G·U wobble pair is exchanged for an A·U Watson–Crick pair at the bottom of the stem (Fig. 2 *B* and *C*). It is noteworthy that two different haplotype blocks exist in *D. pseudoobscura* populations. These haplotype blocks are likely to predate the species split.

Besides ha1 and ha2, three other haplotypes are present in the sample. Sequence comparison of ha1 and ha2 (see Fig. 2*A*) reveals that the most frequent one of these haplotypes (26 of 99) is composed of the 5′ end of ha1 and the 3′ end of ha2, with an obvious break (in sequence) between position 1467 and 1471. This suggests that this haplotype is a recombinant. However, a reciprocal recombinational type has not been found in the sample. The remaining two haplotypes (frequencies: 8/99 and 2/99, respectively) seem to be also recombinants, composed of the 5′ end of ha1 and the 3′ end of ha2, but with breaks in different positions. In both cases, however, reciprocal haplotypes are not present in the sample. The 16 DNA polymorphisms that distinguish the five haplotypes in nt 1464–1500 map to the lower stem of the proposed secondary structure and to the bulge loop; none to the more conserved upper part of the structure (Fig. 2*C*). The observed significant disequilibria are formed between polymorphisms in the 5′ part of the lower helix and the bulge loop (nt 1464–1473), and between the bulge loop and the 3′ part of the lower helix (nt 1473–1500). In contrast, no significant disequilibria have been detected between polymorphisms in the 5′ and 3′ parts of the lower helix (i.e., between polymorphisms with the largest physical distance). This lack of significant disequilibrium between these longer-range pairs of polymorphisms is consistent with the high frequency of putative recombinants (discussed above). It may also indicate that the pairing of the putative lower stem of the intron 2 structure is weak (if it exists at all).

Intron 1. We reanalyzed the hairpin structure in intron 1 found previously by phylogenetic comparison (19). Of the 10 species used in ref. 19, we consider here only 6 because these could be aligned over the entire intron 1 (Fig. 3). These include the species used for the adult intron (except *D. lebanonensis*) and one species from the subgenus *Drosophila*. Among these six sequences, two covariations were found in the boxed regions. The LRT produced $\lambda = 4.01$ and LRT statistic = 25.21 (Table 1). None of the 50 simulation runs resulted in higher LRTs for the permuted sequences (the maximum was 19.45). These results strongly support the phylogenetically predicted pairing region of intron 1. The structure is conserved across all *Drosophila* species compared, including those from the subgenus *Drosophila* and *D. lebanonensis* (19). The hairpins of the *obscura* group species, *D. pseudoobscura* and *D. ambigua*, appear to be particularly stable. Both consist of a single stem with 9 consecutive base pairs (see ref. 19).

```
con                          TTCCAT                ATGGAA
mel gtaactatgcgatg--cccaca-gg CTCCAT gcag------cg ATGGAG g-ttaa-tctcgtgtat--tcaatcc---tag
tei gtaactatgcgctg--cacaca-gg aTCCAT Ttcg-------G ATGGAg g-ttaa a-ctcgagtat--tccatcc---tag
ere gtaag--ggcgatgtgcgcacatgc aTCcAT tg---------g ATtGAg g-ttaa atttcgtgtta--tccatcc---tag
psu gtaag--agtga-------acg-aA TTCCAT GGagt-----CT ATGGAA Tcctaa atttaaaaat---tcatttattttag
amb gtaag---gcga------catc-tA TTCCAT AGagtcctaaCT ATGGAA Tcctaa -tcccgtaatt--ccccacca--tag
hyd gtaa----gcga----------gt GTCIGT gtg-------ta ATAGAT ccctaa atataagcttgacttgttcct--cag
```

FIG. 3. Sequence alignment of intron 1. Five species are from the subgenus *Sophophora*: three from the *melanogaster* species group (mel, *D. melanogaster*; tei, *D. teissieri*; ere, *D. erecta*), and two species are from the *obscura* group (psu, *D. pseudoobscura*; amb, *D. ambigua*), and one species is from the subgenus *Drosophila* [hyd, *D. hydei* (*Adh-2*)].

In contrast to the adult intron and intron 2, linkage disequilibria associated with the structure in intron 1 have not been observed in *D. pseudoobscura* populations (12). All nucleotide polymorphisms segregating within the segment where the structure is located are in low frequency. The four polymorphisms that occur within the pairing region have frequencies of <10% in the sample. All four are due to single mutations (without compensation) and should, therefore, have a destabilizing effect on the structure.

DISCUSSION

By using phylogenetic DNA sequence comparison and a likelihood approach, we have inferred pre-mRNA secondary structures in the three intron regions of the *Drosophila Adh* gene. In each intron, we identified one structure. All pairing regions were phylogenetically predicted based on the Woese–Noller criterion of at least two covariations. The predicted helices in the adult intron and intron 1 were strongly supported by the LRT and the simulations. In intron 2, statistical support could only be found for the very short upper stem. The three inferred structures are considerably different in shape, and each one varies among species. The lower stem of the structure in the adult intron and the hairpin in intron 1 are both the largest pairing regions and are most conserved. (In these two cases, our statistical method also gave the strongest support.) In contrast, the lower stem of the intron 2 structure is least conserved, so that only sequences between very closely related species could be lined up in this region. The observed significant linkage disequilibria associated with these structures tend to fall into the less conserved parts of a structure: In the adult intron, five of seven polymorphisms that form significant nonrandom associations and are thought to be under epistatic selection (12) map to the less conserved upper part of the structure (including the bulge loop, the upper stem, and the upper loop); in intron 2, all polymorphisms resulting in significant disequilibria fall into the variable lower part of the structure (including the lower stem and the large bulge loop); and in intron 1, disequilibria associated with this rather conserved hairpin structure have not been observed.

A quantitative evolutionary model that could explain these observations is currently not available. However, our findings are qualitatively consistent with a model of compensatory fitness mutations. This model assumes that mutations disrupting pairing in a secondary structural element are individually deleterious in a functionally important structure; a second compensatory mutation, however, may restabilize the structure and restore fitness. Our observations summarized above indicate that the presence or absence of strong linkage disequilibria in the three regions of secondary structure may be determined largely by selection pressure against single mutations that could destabilize a pairing region. If selection is weak, single mutations within pairing regions can stay in a population long enough and wait for compensatory mutations to occur on the same chromosome. Thus, the majority of nonrandom associations should be found in portions of a structure that are less conserved. This indeed seems to be consistent with the pattern of linkage disequilibria observed in the adult intron and in intron 2. On the other hand, if selection pressure against destabilizing single mutations is strong, then polymorphisms within pairing regions are expected to be eliminated from a population or stay in low frequency, so that the occurrence of compensatory mutations and, hence, linkage disequilibria are less likely. This latter situation may apply to the hairpin structure in intron 1, which consists of a long 9-bp helix without a bulge loop and is energetically very stable (19). All of the polymorphisms associated with this structure are in low frequency and linkage disequilibria have not been found.

Our interpretation of the action of natural selection on the inferred pre-mRNA secondary structures rests on the assumption that other evolutionary forces, in particular recombination, are less important or have similar effects on the three inferred structures. To understand the process of compensatory evolution in more detail, a quantitative model is needed that ties the patterns of interspecific divergence and intraspecific variation together. Kimura (30) discussed the role of compensatory mutations in interspecific divergence, but this work is not particularly useful in understanding the complex pattern of intraspecific polymorphism and linkage disequilibrium observed in the data at hand.

The generality of our results remains to be seen. A similar pattern of linkage disequilibria has been reported for the white locus of *D. melanogaster*. At white clustering of linkage disequilibria also occurs predominantly in introns (10, 11). Although the mechanism for the disequilibria at white is still unknown, these findings raise important questions about the function of secondary structure in introns and the nature of the selective forces causing the observed differences in linkage disequilibrium patterns between intron and exon sequences.

We thank J. Parsch and three anonymous reviewers for comments. This work was supported in part by National Institutes of Health Grant GM 46233 and a Biomedical Research Award from the University of Maryland to W.S. and by National Institutes of Health Grant GM 16250 to S.V.M. Computing support was provided by the Pennsylvania State University Center for Computational Biology.

1. Haldane, J. B. S. (1931) *Proc. Cambridge Philos. Soc.* **27,** 137–142.
2. Wright, S. (1932) *Proc. Sixth Int. Congr. Genet.* **1,** 356–366.
3. Langley, C. H. (1977) in *Measuring Selection in Natural Populations*, eds. Christiansen, F. B. & Fenchel, T. M. (Springer, Berlin), pp. 265–273.
4. Hedrick, P. H., Jain, S. & Holden, L. (1978) *Evol. Biol.* **11,** 101–182.
5. Clegg, M. T. (1978) *Theor. Popul. Biol.* **13,** 1–23.
6. Lewontin, R. C. (1985) *Annu. Rev. Genet.* **19,** 81–102.
7. Zapata, C. & Alvarez, G. (1992) *Evolution* **46,** 1900–1917.
8. Aguadé, M., Miyashita, N. & Langley, C. H. (1989) *Genetics* **122,** 607–615.
9. Macpherson, J. N., Weir, B. S. & Leigh Brown, A. J. (1990) *Genetics* **126,** 121–129.
10. Miyashita, N. & Langley, C. H. (1988) *Genetics* **120,** 199–212.
11. Miyashita, N. T., Aguade, M. & Langley, C. H. (1993) *Genet. Res.* **62,** 101–109.
12. Schaeffer, S. W. & Miller, E. (1993) *Genetics* **135,** 541–552.
13. Schaeffer, S. W. & Miller, E. (1992) *Genetics* **132,** 163–178.
14. Kimura, M. (1956) *Evolution* **10,** 278–287.
15. Lewontin, R. C. (1974) *The Genetic Basis of Evolutionary Change* (Columbia Univ. Press, New York).
16. Hill, W. G. & Robertson, A. (1968) *Theor. Appl. Genet.* **38,** 226–231.
17. Nei, M. & Li, W.-H. (1973) *Genetics* **75,** 213–219.
18. Ohta, T. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 1940–1944.
19. Stephan, W. & Kirby, D. A. (1993) *Genetics* **135,** 97–103.
20. Fox, G. E. & Woese, C. R. (1975) *Nature (London)* **256,** 505–507.
21. Noller, H. F. & Woese, C. R. (1981) *Science* **212,** 403–411.
22. Muse. S. V. (1995) *Genetics* **139,** 1429–1439.
23. James, B. D., Olsen, G. J., Liu, J. & Pace, N. R. (1988) *Cell* **52,** 19–26.
24. Marfany, G. & Gonzalez-Duarte, R. (1991) *J. Mol. Evol.* **32,** 454–462.
25. Juan, E., Papaceit, M. & Quintana, A. (1990) *Nucleic Acids Res.* **18,** 6420.
26. Mount, S. M., Burks, C., Hertz, G., Stormo, G. D., White, O. & Fields, C. (1992) *Nucleic Acids Res.* **20,** 4255–4262.
27. Felsenstein, J. (1981) *J. Mol. Evol.* **17,** 368–376.
28. Devereux, J., Maederli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12,** 384–395.
29 Zuker, M. (1989) *Science* **244,** 48–52.
30. Kimura, M. (1985) *J. Genet.* **64,** 7–19.