Concerted Transpositions of Mobile Genetic Elements Coupled with Fitness Changes in *Drosophila melanogaster*

E. G. Pasyukova,* E. Sp. Belyaeva,* G. L. Kogan,* L. Z. Kaidanov,†
and V. A. Gvozdev*

*Institute of Molecular Genetics, USSR Academy of Sciences; and †Faculty of Genetics, Leningrad State University

In an inbred low-activity (LA) strain of *Drosophila melanogaster* with a low level of fitness and a complex of inadaptive characters, in situ hybridization reveals an invariant pattern of distribution of three copia-like elements (*mdg-1*, *mdg-3*, and *copia*). Rare, spontaneous, multiple transpositions of mobile elements in the LA strain were shown to be coupled with a drastic increase of fitness. A changed pattern of various types of mobile elements was also observed on selecting the LA strain for higher fitness. High-fitness strains show transpositions of mobile elements to definite chromosomal sites ("hot spots"). Concerted changes in the location of three different mobile elements were found to be coupled with an increase of fitness. The *mdg-1* distribution patterns were also examined in two low-fitness strains independently selected from the high-fitness ones. Fitness decrease was accompanied by *mdg-1* excision from the hot spots of their location usually detected in the high-fitness strains. The results suggest the existence of a system of adaptive transpositions of mobile elements that takes part in fitness control.

Introduction

Ever since *Drosophila melanogaster* was found to have mobile genetic elements whose distribution in chromosomes varied not only from stock to stock but in individuals within a stock (Georgiev et al. 1977; Ananiev et al. 1978; Young 1979), it has been clear that the mobile elements constitute a tremendous potential for genome variability. A study of the distribution of mobile elements in natural populations of *D. melanogaster* has revealed a wide range of genome variability provided by the diverse distribution patterns of different families of mobile elements (Montgomery and Langley 1983). Mobile elements are regarded by some authors as "selfish" DNA whose presence does not in any way improve the fitness of individuals and whose expansion in the genome is limited, say, by natural selection (Dover and Doolittle 1980; Orgel et al. 1980). The role of mobile elements is also frequently discussed in the context of their appreciable contribution to mutational variability (Woodruff et al. 1983). On the other hand, one may hypothesize that at least some types of mobile elements are vital functional components of the genome. There are data that mobile elements are not randomly distributed in the genome but have their preferred sites (hot spots) (Gvozdev 1981; Belyaeva et al. 1984). One cannot rule out the possibility that the transpositions of mobile elements to specific chromosome regions may perform certain biological functions. From this point of view, the results reported in this paper...
may indicate the existence of a special system of mobile elements whose transpositions possibly play an essential part in evolution, since the observed transpositions were coupled with an increase of fitness of individuals.

This study examines one class of mobile elements in *D. melanogaster* (Tchurikov et al. 1981; Rubin 1983): copia-like elements or mobile dispersed genes (*mdg*), which are characterized by certain distinctive features of molecular organization.

**Material and Methods**

**Fitness Estimation**

We used the competition index according to Hartl and Haymer (1983) to characterize fitness. The tested strain was placed in a tube with flies carrying marked compound arms of the second pair of autosomes: C(2L)RM, dp; C(2R)RM, px. The progeny of a cross between this strain and the tested strain without structural rearrangements die as a result of a disturbed genic balance. Therefore it can be assumed that these strains develop under the same conditions in reproductive isolation. The ratio of the number of offspring of the tested strain to the total number of offspring (i.e., the competition index) characterizes the fitness of the tested strain. We analyzed 15–20 test tubes (2,000–5,000 offspring) for each strain. The initial ratio of parental individuals in each test tube was five pairs of tested strain to 15 pairs of tester strain.

**The Strains**

The strains used in this study originated from a low-activity (LA) strain of *Drosophila melanogaster* that was obtained as a result of inbreeding and long-term selection for low mating success from the natural population Yessentuki (Kaidanov 1980). The strain acquired a number of maladaptive properties, namely, low viability and fertility, low mobility, and elevated temperature sensitivity. It is characterized by a very low level of fitness according to Hartl and Haymer's test (Belyaeva et al. 1982). For many years the LA strain has been maintained as families obtained from individual pairs. The rare tubes containing larger numbers of mobile flies were discarded from the collection of LA tubes. This phenomenon of the emergence of high-viability flies is considered in the present paper.

High-fitness strains 171, 68, and 6 arose spontaneously from individual pairs of the LA strain. Strains LA+, HA, and LAi+ were obtained from LA as a result of selection for an increased number of abdominal bristles and a higher male mating activity (Kaidanov 1980). High-fitness strains LA+3, 3, 96, 41.2, and 41.1 were obtained from LA through replacing inbreeding by mass breeding (Belyaeva et al. 1982) and selecting test tubes with larger numbers of offspring. Simultaneously, backward selection for low male mating activity and low viability produced two strains with a low level of fitness: LA− from LA+ and HA− from HA (Kaidanov et al. 1983).

The in situ hybridization and labeling of DNA probes have been described before (Belyaeva et al. 1984). We have used *mdg-1* and *mdg-3* clones (Tchurikov et al. 1981) and the *copia* element within Dm5002 (Dunsmuir et al. 1980).

**Results**

The distribution of mobile genes was studied in a low-fitness, inbred LA strain and related high-fitness strains. Hartl’s fitness test (Hartl and Haymer 1983) correlated quite well with the independent determination of such fitness components as viability, mating success, and fertility (Kaidanov 1980; Kaidanov et al. 1983).

Inbred LA individuals showed a nearly invariant distribution of *mdg-1*, *mdg-3*
Table 1  
Location of Mobile Elements in Chromosomes of the LA Strain

<table>
<thead>
<tr>
<th>MOBILE ELEMENT</th>
<th>NO. OF LARVAE</th>
<th>X Chromosome</th>
<th>CHROMOSOME 2</th>
<th>CHROMOSOME 3</th>
<th>CHROMOCENTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>mdg-1</td>
<td>100</td>
<td>19F</td>
<td>23A, 30A, 33C, 34EF, 35DE, 56F, 59CD</td>
<td>75F, 82DE, 84A, 88E, 96BC, 98BC, 99A</td>
<td>-</td>
</tr>
<tr>
<td>mdg-3</td>
<td>10</td>
<td>7A, 7E</td>
<td>34D, 34E, 41AD</td>
<td>65A, 67BC, 90AB, 99D, 99F, 100F</td>
<td>-</td>
</tr>
<tr>
<td>copia</td>
<td>10</td>
<td>5A*, 18AC</td>
<td>23A, 26C, 33A, 34D, 36A, 39D, 42B, 47D, 53E</td>
<td>64E, 66BC, 71C, 73C, 80, 82C, 87EF, 94EF, 96B, 98C, 99B</td>
<td>-</td>
</tr>
<tr>
<td>mdg-4 (gypsy)</td>
<td>5</td>
<td>52B</td>
<td></td>
<td></td>
<td>+</td>
</tr>
</tbody>
</table>

* Hybridization in the 5A region is from the single-copy DNA sequence within the recombinant plasmid.
(Belyaeva et al. 1982), and copia (table 1), located according to the Bridges cytological map. Earlier, a change of the mdg-1 distribution was described on selection of the LA strain for higher fitness (Gvozdev et al. 1981). Mass breeding of the LA stock also yielded strains with an increased fitness and changed mdg-1 pattern (Belyaeva et al. 1982).

The LA strain shows remarkable stability with respect to the mdg-1 distribution (table 2), but during long-term selection and mass breeding it is difficult to exclude a selection of very rare individuals with a changed mdg pattern: they are the ones that presumably have an increased fitness. To reliably demonstrate the existence of transpositions attending the observed phenotypic changes in LA individuals, it was necessary to analyze the distribution of mdg-1 in the progeny of one pair of flies with the known characteristic LA pattern of mdg.

The original LA families were obtained through 19 LA × 18 LA crosses. The mdg-1 distribution was determined in some of the larvae (N = 5–10 individuals). All tested larvae had the same characteristic LA pattern (table 1). The rest of the progeny were used to obtain the next generation from individual pairs. All subsequent generations were also obtained by individual sib crosses. Test tubes with a large number of abundant and mobile offspring were selected. Three families were found (experiments were performed with considerable time intervals) to have the expected phenotypic changes in the fourth (strain 171), sixth (strain 6), and seventh (strain 68) generations. A total of 8 × 10³–1 × 10⁴ individual families were analyzed in each experiment. A rough estimation of the frequency of occurrence of high-fitness families must be ~3 × 10⁻⁴. The very low competition index of the LA strain (0.04 ± 0.006) increased 15–20-fold in strains 171, 68, and 6 (to 0.70 ± 0.07, 0.76 ± 0.04, and 0.62 ± 0.03, respectively). The distribution of mdg-1, mdg-3, and copia also changed in these strains compared with the original LA strain. After detection of fitness increase, the locations of mobile elements were determined in the first and several subsequent generations; no drastic changes in the newly established mobile-element patterns were observed when these strains were reinvestigated after a year. The progeny of the other individual families that retained the phenotype of the LA strain (i.e., small number of offspring,

Table 2
Changes in Number of the mdg-1 and mdg-3 Sites in the LA and High-Fitness Strains

<table>
<thead>
<tr>
<th>STRAIN</th>
<th>YEAR</th>
<th>NO. OF SITES</th>
<th>LA Sites Disappearing</th>
<th>New Sites Appearing</th>
<th>NO. OF LARVAE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>mdg-1</td>
<td>mdg-3</td>
<td>mdg-1</td>
<td>mdg-3</td>
</tr>
<tr>
<td>LA</td>
<td>1980</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2*</td>
</tr>
<tr>
<td></td>
<td>1983</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3*</td>
</tr>
<tr>
<td></td>
<td>1984</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3*</td>
</tr>
<tr>
<td>68</td>
<td>1980</td>
<td>9</td>
<td>10</td>
<td>45</td>
<td>22</td>
</tr>
<tr>
<td>171</td>
<td>1980</td>
<td>12</td>
<td>10</td>
<td>29</td>
<td>15</td>
</tr>
<tr>
<td>1983</td>
<td>0</td>
<td>3</td>
<td>18</td>
<td>18</td>
<td>13</td>
</tr>
</tbody>
</table>

Note.—Changes in LA strain are relative to the data obtained in 1978.
* Sites showing weak hybridization.
small size, and low mobility of flies) and a low level of fitness were found to retain
the original LA pattern of mobile elements (with exceptions concerning the weak spots
of hybridization; table 2).

Since the chromosomes of the initial parental flies had the characteristic LA
pattern of \textit{mdg-1}, the change observed in strains 6, 68, and 171 obtained from one
pair of individuals can only be the result of \textit{mdg} transpositions that occurred during
the experiment.

Figure 1 shows the results of the distribution analysis of the mobile elements in
chromosome 2. All sites detected are shown. We analyzed 13–20 larvae of each strain.
The individuals belonging to one strain show a certain degree of heterogeneity in the
\textit{mdg} distribution, as can be demonstrated in the case of strain 6: among the 13 tested
larvae of strain 6, 10 carried LA-specific \textit{mdg-1} sites in the 23A region, 11 in 30A, 7
in 33C, 8 in 34EF, and 10 in 35CD; all individuals conserved \textit{mdg-1} copies in 56F
and 59CD; 8 carried new \textit{mdg-I} copies in 37CD, 9 in 39CD, 11 in 41 and 42A, 4 in
47D, 3 in 52E and 53C, 4 in 57B, and 7 in 55D. The high-fitness strains analyzed
showed a similar degree of heterogeneity with regard to the other mobile elements
data not shown).

The pattern of mobile elements characteristic of the LA strain (table 1), which
is composed of the \textit{mdg-1}, \textit{mdg-3}, and \textit{copia} sites (black squares, fig. 1), is almost fully
preserved in strain 6, except for the fact that \textit{copia} disappears from 36A and 47D.
Against the background of this characteristic LA pattern, a number of new sites appear.
Notice the joint distribution of the three mobile elements in strains 6 and 171 (asterisks,
fig. 1). The joint distribution of the mobile-element patterns in strains 171 and 68
also reveals their similarity. The similarity is emphasized by the circles in figure 1 that
mark 13 shared sites of different mobile elements.

Note that all three strains (6, 68, and 171) retain an unchanged location of
\textit{mdg-4} (gypsy) in the 52B region of chromosome 2, a feature that is characteristic of
the original LA strain (table 1). Being a representative of \textit{copia}-like elements (see
Djumagaliev et al. 1983; Freund and Meselson 1984), \textit{mdg-4} (gypsy) usually occurs
in from three to five variable sites in the polytene chromosomes of other \textit{Drosophila
melanogaster} strains (Ananiev et al. 1984). This suggests that these high-fitness strains
did not arise by contamination from other \textit{D. melanogaster} strains.

The results demonstrate a similarity in the transposition pattern of mobile ele-
ments in three high-fitness strains obtained from individual pairs. Figure 1 also shows
the results of the analysis of strains 3 and 96, with a sharply increased fitness (the
competition indices are 0.77 ± 0.03 and 0.75 ± 0.04, respectively) revealed during
mass breeding of LA flies. As in strain 6, the pattern of mobile elements characteristic
of the low-fitness strain (black squares, fig. 1) is largely preserved, but new sites also
appear (underlined slanting strokes, fig. 1). Once again, there is a great similarity
between strains 3 and 96 with respect to the complex pattern of mobile elements in
eight sites of chromosome 2. The last line in figure 1 shows the distribution of mobile
elements in a high-fitness HA strain (competition index 0.43 ± 0.01) obtained from
LA by long-term selection. Again, many new sites (slanting strokes and triangles, fig.
1) appear alongside “traces” of the characteristic LA pattern; the 17 sites marked by
triangles point to the similarity of the mobile-element patterns in strains 3 and HA.

These results display instances of concerted transpositions of mobile elements.
Nonrandom transpositions to specific sites with a concomitant increase in fitness are
demonstrated by the overall analysis of all the high-fitness derivatives of the LA strain.
The analysis of 11 such strains (171, 68, 6, LA⁺, LA, LA₁⁺, LA₂⁺, 3, 96, 41.2, and
411; see Material and Methods) is summarized in figure 2. Arrows indicate mobile-element sites occurring both in LA and in derived strains. For example, most of the high-fitness strains retain, at least in some individuals, mdg-1 at 30A, 34EF, and 56F.

**Subdivisions according to cytological map**

![Diagram showing mobile element sites](image)

Fig. 1.—Patterns of mobile elements (mdg-1, copia, and mdg-3) in chromosome 2 in the original low-fitness LA strain and in related high-fitness strains; black squares mark the characteristic LA sites; any box that is not empty indicates that the element noted at the left of the chart was observed at the chromosomal position noted at the top of the chart. This is simply a slash (/) in many cases. Where the same element has independently moved to the same position in two (or more) separate strains, the slash has been replaced by other symbols to emphasize a particular pair of strains as follows: (•: 6, 171; O: 68, 171; Δ 3, HA). For the pair (3, 96) the slash is not replaced, but a bold horizontal line has been placed at the bottom of the square. Superimposed symbols imply that the element found in that strain is also found in at least two other strains at that position. Thirteen to 20 larvae of each strain were examined. Not all larvae showed all of the indicated elements, but all have most of them.
The copia element is usually retained at 23A, 26C, 39CD, and 42B. Against this background one observes hot spots of mobile-element transpositions in more than five of the 11 high-fitness strains studied. For mdg-1 these are regions 11C and 19B in the X chromosome and regions 22B, 30CD, 36CD, 37CD, 39CD, 41, and 42A in chromosome 2 (marked in fig. 1). Regions 13A and 42B contain hot spots for mdg-3. In five of the 11 cases studied, transpositions of copia to regions 11C, 34F, 35CD, and 57A occurred. If one allows for the fact that in situ hybridization enables approximately 400 mobile-element sites to be revealed in the polytene chromosomes, the analysis of the distribution of three mobile elements over the entire genome of 11 strains (data not shown) demonstrates that the observed distribution is quite different from the theoretical Poisson distribution (table 3).

Interestingly, the frequently occurring regions of mdg-1 and copia transpositions (11C and 57A)—as determined within the range of accuracy of the in situ–hybridization technique—may overlap (fig. 2). In many cases the location of mdg-1 in the LA strain coincides with the copia transposition sites (30A, 34F, 35CD, and 59CD) attending increased fitness. The regions of mdg-1 transpositions may coincide with the copia sites in the LA strain (26C, 33A, and 39CD). The 34D location of mdg-3 in LA coincides with an mdg-1 transposition site. The 42B region in which copia is localized in the LA strain is the hottest spot for mdg-3 transpositions. These results demonstrate not only the nonrandomness of transpositions to specific chromosome sites but the tendency of different mobile elements to be inserted into the same region.
Table 3
The Observed and Expected Distributions of Mobile Elements over All Chromosomes in 11 High-Fitness Strains

<table>
<thead>
<tr>
<th>No. of Hydration Instances in Region</th>
<th>mdg-1 Observed</th>
<th>Random</th>
<th>mdg-3 Observed</th>
<th>Random</th>
<th>copia Observed</th>
<th>Random</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>264</td>
<td>176</td>
<td>274</td>
<td>238</td>
<td>240</td>
<td>197</td>
</tr>
<tr>
<td>1</td>
<td>63</td>
<td>144</td>
<td>79</td>
<td>124</td>
<td>89</td>
<td>140</td>
</tr>
<tr>
<td>2</td>
<td>27</td>
<td>59</td>
<td>30</td>
<td>32</td>
<td>36</td>
<td>50</td>
</tr>
<tr>
<td>3</td>
<td>17</td>
<td>16</td>
<td>9</td>
<td>6</td>
<td>21</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>11</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>10</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>0.5</td>
<td>2</td>
<td>0</td>
<td>3</td>
<td>0.4</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

\[ \chi^2 = 137 \quad \chi^2 = 36.5 \quad \chi^2 = 58.4 \]

\[ p < 0.001 \quad p < 0.001 \quad p < 0.001 \]

Transposition hot spots are also found in chromosome 3 in 11 high-fitness strains. These are regions 75A, 94, and 97A for *mdg-1* (in six or more strains), 69B and 96A for *copia* (in five and six strains, respectively), and 64A for *mdg-3* (five strains). Here again, though to a lesser extent than in chromosome 2, one can observe co-occurring transpositions of different mobile elements coupled with the increase of fitness. For example, both in strain 6 and in strain 41.1 simultaneous transpositions of *mdg-1* in 75A and 97A and of *copia* in 65E, 70A, 72A, and 96A were detected.

The preliminary attempt to correlate the fitness level and the distribution of mobile elements can be made by examining the pattern of *mdg-1* in 14 related strains derived from LA, including, in addition to the above-mentioned 11 high-fitness strains, two strains with a very low level of fitness, a level that comes close to that of LA if assessed by the competition index (table 4). The low-fitness LA- and HA- strains were obtained by reverse selection from LA+ and HA, respectively (see Material and Methods). In our analysis of the distribution of mobile elements among these strains we consider only the hot spots (detected in more than five strains) of the high-fitness strains' transpositions. There are 11 such spots in the genome (tables 2, 3). The sites characteristic of the LA strain are not taken into consideration: we assume that the LA chromosomes do not contain any hot spots for *mdg-1*. Table 3 shows that, as a rule, the acquisition of a high level of fitness is accompanied by the appearance of 6–10 new *mdg-1* sites in the hot spots (with the exception of strain 41.1, which displays only three new hot spots for *mdg-1*). On reverse selection to low fitness, the majority of *mdg-1* copies disappear from the hot spots, whose number dwindles to two.

These results suggest that the distribution of mobile elements in hot spots is related to the level of fitness.

The analysis of other mobile elements in the LA- and HA- strains does not so far reveal any such simple correlation as the comparison of the *mdg-1* pattern and
Competition Index (Fitness) and the Number of Hot Spots for *mdg-1* in Related Strains

<table>
<thead>
<tr>
<th>STRAIN</th>
<th>COMPETITION INDEX</th>
<th>HOT SPOTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>41.2</td>
<td>0.93 ± 0.06</td>
<td>+ + + + + + + + + + 10</td>
</tr>
<tr>
<td>LA^+</td>
<td>0.91 ± 0.04</td>
<td>+ + - + + + + + - + 8</td>
</tr>
<tr>
<td>3</td>
<td>0.77 ± 0.03</td>
<td>+ + - + + + + - + - + 8</td>
</tr>
<tr>
<td>LA^2+</td>
<td>0.77 ± 0.03</td>
<td>+ + + + + + - + - - + 8</td>
</tr>
<tr>
<td>68</td>
<td>0.76 ± 0.04</td>
<td>+ + + - - + + + + - + 6</td>
</tr>
<tr>
<td>96</td>
<td>0.75 ± 0.04</td>
<td>+ + + - + + + - + + - + 9</td>
</tr>
<tr>
<td>41.1</td>
<td>0.73 ± 0.03</td>
<td>+ - - - - - + - - + - 3</td>
</tr>
<tr>
<td>171</td>
<td>0.70 ± 0.07</td>
<td>+ - - + + + + + - + - 9</td>
</tr>
<tr>
<td>LA^1+</td>
<td>0.65 ± 0.045</td>
<td>+ + + + + + - - - - + 8</td>
</tr>
<tr>
<td>6</td>
<td>0.62 ± 0.03</td>
<td>+ - - - - - + + + - + 7</td>
</tr>
<tr>
<td>HA</td>
<td>0.43 ± 0.01</td>
<td>- + + + + + - + + - + 7</td>
</tr>
<tr>
<td>HA^-</td>
<td>0.05 ± 0.007</td>
<td>- - - - - - - + - + + 2</td>
</tr>
<tr>
<td>LA^-</td>
<td>0.04 ± 0.015</td>
<td>- - - - - - - + - + - + 2</td>
</tr>
<tr>
<td>LA</td>
<td>0.04 ± 0.006</td>
<td>- - - - - - - - - - - 0</td>
</tr>
</tbody>
</table>

The suddenly increased fitness in LA stock may be suspected as arising from contamination. There are several reasons that argue against this supposition. For example, it is impossible to explain by contamination the emergence of high-fitness strain 6, which conserves the majority of molecular markers peculiar to the progenitor LA strain. The similarities of the new patterns of mobile genes in the independently obtained 68 and 171 strains are also regarded as evidence against contamination.

Contamination as the cause of the drastic increase in fitness may be rejected when transpositions of mobile elements are observed against the background of sites preserved from the parental strain. In these experiments pieces of the same gland were used for in situ hybridization with *mdg-1*, *mdg-3*, and *copia*. The characteristic sets of mobile elements peculiar to the progenitor LA stock remained unchanged in several parts of the high-fitness genome and, at the same time, several mobile elements newly appeared in these regions. Several such examples are presented in table 5. In strain 41.1 the typical LA pattern of *mdg-1* is conserved in the 2L chromosome, although a new site of *copia* at 32A and four new sites of *mdg-3* were detected. In strain 6 (larva 1) three new sites of *copia* element appeared against the peculiar LA pattern, including the 53E *copia* site and two *mdg-1* sites, 56F and 59CD. In strain 6 (larva 2) two *copia* sites and a single *mdg-3* site were detected in the pericentromeric region of the 2R chromosome marked by LA sites at the 41AD (*mdg-3*) and 42B (*copia*) regions. Finally, at the top of the 3R chromosome in strain LA^1+, LA-specific sites at the 98BC (*mdg-1* and *copia*) and at the 100F (*mdg-3*) flank the newly appearing sites at 100A (*mdg-1*) and 99D (*copia*).

Situated in different genomic regions, these molecular markers (only part of which are presented in table 5) argue strongly for the origination of high-fitness stocks from the progenitor LA stock.
### Table 5
**Changed Patterns of Different Mobile-Element Locations Detected in Single Larvae**

<table>
<thead>
<tr>
<th>Strain 41.1 (chromosome 2L):</th>
<th>Location(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mdg-1</td>
<td>30A</td>
</tr>
<tr>
<td>copia</td>
<td></td>
</tr>
<tr>
<td>mdg-3</td>
<td>30A 30CD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strain 6 (chromosome 2R):</th>
<th>Location(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Larva 1:</td>
<td></td>
</tr>
<tr>
<td>mdg-1</td>
<td></td>
</tr>
<tr>
<td>copia</td>
<td>53E 54D</td>
</tr>
<tr>
<td>mdg-3</td>
<td>(no sites)</td>
</tr>
<tr>
<td>Larva 2:</td>
<td></td>
</tr>
<tr>
<td>copia</td>
<td>41F</td>
</tr>
<tr>
<td>mdg-3</td>
<td>41AD</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strain LA_{1\ast} (tip of chromosome 3R):</th>
<th>Location(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>mdg-1</td>
<td>98BC</td>
</tr>
<tr>
<td>copia</td>
<td>98C 99D</td>
</tr>
<tr>
<td>mdg-3</td>
<td></td>
</tr>
</tbody>
</table>

**Note:** Boxed regions designate the LA-specific sites of mobile-element locations.
Table 5 also supports the notion of co-occurring behavior of different mobile elements as partners in the transposition event (this feature of transposition process has been mentioned in Results; also see fig. 2). Table 5 shows that a new mdg-3 copy was detected in strain 41.1 in the 30A region occupied by mdg-1; copia appeared in strain 6 in the 56F region occupied by mdg-1; mdg-3 transposed to the 42B region occupied by copia. The co-occurrence of different mobile elements in these sites that was revealed by in situ hybridization is corroborated by the isolation of genomic cosmid clones carrying different mobile elements; for example mdg-1 and mdg-3 co-occur in the same DNA fragment adjacent to unique DNA from the 34E region (Yu. Ya. Shevelev, personal communication).

Discussion

It has generally been accepted that fitness variability is largely dependent on the mutation rate of the components of a polygenic system and, perhaps to an even greater extent, on effective recombination of different polygenes involving crossing-over. Hence change of fitness has been regarded as a fairly slow multistage process involving a system of individual genes (Dobzhansky 1970; Crow and Simmons 1983).

In the present work we studied a low-fitness LA strain that displayed episodes of sharp spontaneous increase of fitness. This was a good selective system that made it possible to detect rare events leading to increased fitness. These events were always attended by abrupt transpositions of a number of mobile elements. These observations suggest a basically new molecular mechanism that permits a rapid increase of fitness through multiple nonrandom transpositions of mobile elements into specific genome targets in a single generation. The transpositions described above may be the cause of the fitness change; we suppose that they change the level of expression for a whole number of genes. Since the transpositions occur simultaneously, they may switch the activity of a number of genes in a prompt and coordinated fashion. This may lead to saltatory fitness changes, which may be critical to the evolutionary fate of a population.

Looking at various descriptions of the Drosophila fitness dynamics, one notices that fitness may change very rapidly (Hartl and Jungen 1979; Templeton 1979). It is difficult to account for such rapid changes either by the mutation rate of individual genes or by a combination of them leading to favorable sets of polygenes. A possible explanation of the phenomenon may involve transpositions of mobile elements. Then the following features of the transposition process would be noted.

In D. melanogaster mobile elements can be transposed in germline cells in the process of cell divisions during gamete maturation. This can be inferred from the analysis of unstable mutations that are caused by the transposable elements (Green 1980). If this is true in our case, then a new pattern of mdg will characterize a cluster of newly forming gametes, so that the efficiency of directed transpositions will be especially high, since they will at once affect a great number of individuals in the population.

It seems important to emphasize that the observed mdg transpositions take place in individuals of a "pure line." All of them (N > 100 larvae) had the same distribution of mdg-1. Hence the transpositions of mobile elements are the molecular basis of considerable genetic heterogeneity of gametes in pure-line individuals. These results throw new light on the fate of pure lines and on the causes of the genetic heterogeneity that they often display in experiments (Wright 1977).

The transpositions in the LA strain that are coupled by a marked biological effect may be considered in terms of the well-known concept of genetic homeostasis (Lerner 1954). According to this concept, long-term selection, which took place in the case of
the LA strain, may well lead to a considerable loss of fitness because of a disturbance of the optimal balance of putative polygenes that determine fitness components. It is presumed that when selection has to stop because of the drop in viability, genetic homeostatic mechanisms become activated and lead to a recovery of a high level of fitness. So far one can only speculate about the molecular mechanisms of these complex processes, but one certainly cannot rule out the involvement of \textit{mdg} transposition in the recovery of a high level of fitness.

The results reported in this paper may indicate, for the first time, specific molecular mechanisms that provide saltatory events that play an important part in producing adaptive changes—and hence that are important in the evolution of populations. The transpositions of mobile elements described in this paper have much in common with the genome rearrangements postulated by Goldschmidt, an opponent of gradualism and an active advocate of the role of saltatory changes in evolution (Goldschmidt 1944). According to Goldschmidt, "systemic mutations" leading to significant phenotypic results arise from changes in the mutual disposition of genes (chromosomal repatterning), which give rise to new kinds of gene interactions as a result of position effects. Goldschmidt contrasted such putative mutation systems, occurring as a result of single events but leading to strong phenotypic effects, to mutations in individual genes whose phenotypic effects may be comparatively weak. There was a time when Goldschmidt's concepts were criticized as being contradictory to gene theory and the Darwinian theory of evolution. However, now we can see that massive rearrangements, which Goldschmidt considered to be the motive force of macroevolution, do take place but are associated, at least in our system of \textit{D. melanogaster} strains, with microevolutionary changes.

The transpositions described in this paper can be regarded as a reflection of an effective \textit{system of adaptive transpositions} that alters the functional organization of the regions containing mobile elements. The system of adaptive transpositions may create a favorable genetic background for the emergence and establishment of new features caused by mutations in other individual genes. The saltatory increase of fitness in individuals carrying harmful mutations will prevent such individuals from being eliminated and will ensure the coadaptation of the newly emerging variants of genetic interactions.

It is too early to discuss the molecular mechanisms providing the change of the complex set of characters that is fitness, assuming that it may depend on the system of adaptive transpositions of mobile genetic elements. The study of this system of transpositions has only just begun. Certain results (Gvozdev 1981; Belyaeva et al. 1984) make it possible to characterize the hot spots for \textit{mdg-1} as regions of intercalary heterochromatin that have distinctive features of replication and molecular organization. A study of cloned \textit{D. melanogaster} genome regions containing \textit{mdg-1} shows them to contain clustered copies of different repeated mobile elements (Balakireva et al. 1984). These results are consistent with the findings of the present study, which has revealed the tendency of different mobile elements to locate into the same chromosome region. Every such region probably contains mobile elements reshuffled in a certain way. It seems that the increase of fitness depends on the appearance of \textit{mdg} in new regions rather than on their disappearance from the characteristic LA sites, for in some cases abrupt transpositions and a dramatic increase of fitness were observed while practically all the old \textit{mdg-1} sites were preserved. It cannot be ruled out that the transcription intensity of the \textit{mdg} themselves, which may also be a selective factor, sharply changes as a result of the alteration of their immediate environment. According to a more widespread hypothesis, however, \textit{mdg-1} can activate the transcription of
adjacent genes if the transpositions produce new promoter sequences or combinations of sequences that enhance transcription in some way (Di Nocera and Dawid 1983; Steller and Pirrotta 1984). Still, it would be premature to assume that this mechanism of gene activation takes place in the LA strain that undergoes a sharp change in fitness level.

Since the mdg transpositions may be ordered, they can be presumed to cause the change of activity of a number of genes, i.e., produce systemic regulation. Indeed, the transpositions of mobile elements to different regions may create a coordinated system of genes, as has been demonstrated for yeast (Errede et al. 1981).

The results considered above suggest a new approach to the analysis of fitness, an essential biological characteristic of individuals. It seems that either the level of fitness or its components may be analyzed on the basis of the number and pattern of mobile genetic elements in the hot spots of their locations. The regions of hot spots of mobile elements can be compared to polygenic loci, which seem to interact and contribute to the determination of selective value. At least this proposition does not contradict the accumulated data on the genetic analysis of fitness and its components. At the same time we are obviously dealing with a very complicated problem, and we should like to say at once that the extension to natural populations of this simplified method of analysis that we have applied to a highly specialized artificial system of laboratory strains would certainly pose formidable difficulties and might well prove impossible.

LITERATURE CITED


KEN W. JONES, reviewing editor

Received May 6, 1985; revision received September 25, 1985.
We report the nucleotide sequence of a cloned cDNA, pMTS-3, that contains a 1-kb insert corresponding to mouse thymidylate synthase (E.C. 2.1.1.45). The open reading frame of 921 nucleotides from the first AUG to the termination codon specifies a protein with a molecular mass of 34,962 daltons. The predicted amino acid sequence is 90% identical with that of the human enzyme. The mouse sequence also has an extremely high degree of similarity (as much as 55% identity) with prokaryotic thymidylate synthase sequences, indicating that thymidylate synthase is among the most highly conserved proteins studied to date. The similarity is especially pronounced (as much as 80% identity) in the 44-amino-acid region encompassing the binding site for deoxyuridylic acid. The cDNA sequence also suggests that mouse thymidylate synthase mRNA lacks a 3' untranslated region, since the termination codon, UAA, is followed immediately by a poly(A) segment.

Introduction

Thymidylate synthase (E.C. 2.1.1.45) converts dUMP to dTMP by reductive transfer of a methyl group from 5,10-methylenetetrahydrofolate. The enzyme is necessary for the de novo synthesis of TMP and is essential in rapidly proliferating cells. For this reason, it is an important target enzyme for chemotherapeutic drugs (Hartmann and Heidelberger 1961; Danenberg 1977). Thymidylate synthase activity is much greater in exponentially growing cells than in nondividing cells (Maley and Maley 1960; Conrad and Ruddle 1972; Navalgund et al. 1980; Rode et al. 1980) and is synthesized primarily, if not exclusively, during the S phase of the cell cycle (Storms et al. 1984; Jenh et al. 1985a).

To facilitate the analysis of thymidylate synthase and its regulation, an FdUrd-resistant mouse 3T6 cell line (LU3-7) was isolated. The LU3-7 cells overproduce thymidylate synthase by a factor of 50–100 compared with the parental cells (Rossana et al. 1982). By several different criteria, the overproduced enzyme appears to be the same as the wild-type enzyme. However, minor differences cannot be ruled out at present. The overproduction of thymidylate synthase in LU3-7 cells is due to a corresponding increase in the amount of thymidylate synthase mRNA (Geyer and Johnson 1984) and the number of copies of the thymidylate synthase gene (Jenh et al. 1985a). Northern and Southern blot analyses indicate that the size of the mRNA and the physical map of the gene are the same in the overproducing and parental cell.
lines. Since thymidylate synthase gene expression is regulated in LU3-7 cells in a manner similar to that in 3T6 cells (Navalgund et al. 1980; Jenh et al. 1985c), this cell line has been used as a convenient model system for investigating the content and metabolism of thymidylate synthase (Jenh et al. 1985c) and its mRNA (Jenh et al. 1985b).

Previously, we have isolated a number of cDNA clones corresponding to thymidylate synthase mRNA from the overproducing cell line (Geyer and Johnson 1984). The cDNAs were inserted into pBR322 at the PstI site using the poly(dG)/poly(dC) tailing procedure. Several cDNA inserts appeared to be of sufficient length to encode the entire protein. In this paper, we report the DNA sequence of one of these thymidylate synthase cDNA plasmids, pMTS-3. Analysis of this sequence revealed a striking degree of similarity with previously reported sequences of prokaryotic and human thymidylate synthase as well as the apparent absence of a 3' untranslated region on the thymidylate synthase mRNA.

Material and Methods

Strains, Plasmids, and Reagents

The cDNA plasmids pMTS-3 and pMTS-4 (Geyer and Johnson 1984) were maintained in E. coli HB101. M13 derivatives were maintained in JM101. M13 sequencing kits were purchased from New England Biolabs. Restriction endonucleases were obtained from Bethesda Research Laboratories and International Biotechnologies, Inc. Calf intestinal alkaline phosphatase was obtained from Boehringer Mannheim. T4 polynucleotide kinase was obtained from P-L Biochemicals and Amersham. Isotopes were obtained from New England Nuclear.

DNA Sequence Analysis

Sequence analysis of 5' end-labeled fragments was carried out essentially as described by Maxam and Gilbert (1977) using reactions specific for G, G + A, A + C, C + T, and G bases. The nucleotide sequence was confirmed by the dideoxy chain-termination method of Sanger et al. (1977). Restriction fragments of pMTS-3 or pMTS-4 were subcloned into M13 (mp8, mp9, mp18, or mp19), and the single-stranded viral DNA was used as a template for sequence analysis. Sequencing gels were 6%, 8%, or 20% polyacrylamide gels with an acrylamide:bis acrylamide ratio of 19:1. The gels contained 47% urea and 100 mM Tris-borate, pH 8.3, 2 mM EDTA and were run as described by Maxam and Gilbert (1977) and Sanger et al. (1977). Sequences were analyzed using the DNASTAR computer program (provided by F. Blattner and J. Schroeder, University of Wisconsin—Madison).

Results and Discussion

Sequence Determination

The longest of the cDNA clones obtained by Geyer and Johnson (1984), pMTS-4, had a cDNA insert of ~1.1 kb. However, preliminary sequence analysis revealed that 130 nucleotides at the 5' end of pMTS-4 were an inverted repeat of the sequence at the 3' end of pMTS-4 (data not shown). This was most likely the result of a reverse transcriptase cloning artifact. The same type of artifact was observed previously for other cloned cDNAs (e.g., Weaver et al. 1981). For this reason, we focused our analysis on pMTS-3, which had a cDNA insert of ~1.0 kb. The cDNA sequences of pMTS-3 and pMTS-4 are the same except for the 5' end.
Figure 1 shows the restriction map and the strategy for DNA sequence determination of pMTS-3. The nucleotide sequence of the coding strand of the insert DNA is shown in figure 2. Computer analysis of all six possible reading frames revealed a single long open reading frame, which is shown below the nucleotide sequence. The first ATG start codon is shown at position 1. The TAA stop codon is indicated by asterisks below the codon. The stop codon is followed immediately by a stretch of A nucleotides, which probably represents the poly(A) tail of the mRNA (see below). The 5' and 3' ends of the cDNA insert are flanked by oligo(dG) and oligo(dC), respectively, which were added during the cloning of the cDNA (Geyer and Johnson 1984).

The 921-nucleotide sequence from the first ATG to the TAA stop codon predicts a protein with 307 amino acids and a molecular mass of 34,962 daltons. This is slightly lower than the molecular mass of mouse thymidylate synthase protein as estimated by sodium dodecyl sulfate–polyacrylamide gel electrophoresis (38,000 daltons) (Rode et al. 1979; Rossana et al. 1982) but similar to that for the human enzyme (35,706 daltons) as predicted by the cDNA sequence (Takeishi et al. 1985).

At present we cannot be certain that the cDNA represents the entire coding region of the enzyme. The cDNA appears to be complete at the 3' end, since poly(A) is present. However, it is possible that the cDNA does not extend to the initiation codon, since the open reading frame continues upstream of the first AUG codon. Thus a small portion of the amino acid sequence at the amino terminal region of the enzyme may be missing. Several attempts were made to determine the amino acid sequence at the amino terminus of mouse thymidylate synthase. Unfortunately, these attempts were unsuccessful, suggesting that the terminal amino acid of the mouse enzyme may be blocked (J. Freisheim and L. F. Johnson, unpublished data). However, the similarity between the mouse and human thymidylate synthase sequences at the amino terminus of the human protein (see below) does suggest that the entire mouse encoding region is represented.

Conservation of Amino Acid Sequences

In figure 3, the predicted amino acid sequence of mouse thymidylate synthase is compared with the sequences of thymidylate synthase from human (Takeishi et al. 1985).

![Restriction map and sequencing strategy for pMTS-3](image)

**FIG. 1.—Restriction map and sequencing strategy for pMTS-3.** For clarity, only restriction enzymes with 6-base recognition sequences are shown on the map. Restriction sites are represented as follows: B, BamHI; C, ClaI; G, BglII; P, PstI; S, SphI. The arrows indicate the direction and extent to which the sequences were determined. Sequences determined by the Maxam and Gilbert (1977) procedure are indicated below the restriction map, while sequences determined by the M-13/Sanger procedure are shown above the map. The thick lines at both ends of the cDNA insert represent pBR322. Sequences were read across all restriction sites. Approximately 80% of the sequence was determined on both DNA strands.
FIG. 2.—Nucleotide sequence and predicted amino acid sequence of the cDNA insert of pMTS-3. Numbering of both sequences begins at the first AUG codon. The termination codon TAA is indicated with asterisks. The poly(C) and poly(G) at the ends of the sequence were added during the cloning of the cDNA and are immediately adjacent to the PstI sites (not shown) that flank the cDNA insert.
FIG. 3.—Comparison of amino acid sequences of mammalian and prokaryotic thymidylate synthases. The published sequences of human (H), *L. casei* (L), *E. coli* (E), T4 phage (T), and the φ31 phage of *Bacillus subtilis* (P) are aligned with the amino acid sequence of the mouse (M) enzyme from fig. 2. Alignment was performed by computer analysis as well as by visual inspection. Gaps (blank spaces) were introduced to increase identities. Dots indicate that the amino acid at that position is the same as that in the mouse sequence. The position of the amino acid at the right end of each row is indicated by the numbers to the right of the sequences. Region 1 encompasses the folate binding site, region 2 the dUMP binding site (cys at position 189 of the mouse sequence).
1985), *Lactobacillus casei* (Maley et al. 1979), *E. coli* (Belfort et al. 1983), T4 phage (Chu et al. 1984), and the \( \phi 3T \) phage of *Bacillus subtilis* (Kenny et al. 1985). The sequences are aligned so as to increase the regions of identity. The percentages of amino acids that are identical are shown in table 1. As expected, the mouse and human sequences show the highest overall degree of similarity, 90%. The region of lowest similarity is the amino terminal portion of the protein, at amino acids 1-42 of the mouse sequence. If this region is excluded, the mouse and human protein sequences are \( \sim 95\% \) identical.

The mouse and bacterial protein sequences are also quite similar. The overall extent of conservation between the mouse sequence and the *E. coli* sequence (55\%) is almost as great as that between *L. casei* and *E. coli* (60\%). Even the phage and mouse sequences are 40\%-47\% identical. This striking similarity between mammalian and bacterial thymidylate synthase sequences was noted previously by Takeishi et al. (1985) in their analysis of the human sequence. Table 1 shows that region 1 (amino acids 44-70 of the *L. casei* sequence), corresponding to the folate binding site (Maley et al. 1982), is conserved to approximately the same extent as the entire protein. The most conserved domain of the protein is region 2, comprising the 44 amino acids (193-236 of the *L. casei* sequence) surrounding the dUMP binding site (Chu et al. 1984). In this region, the mouse and bacterial sequences are \( \sim 80\% \) identical.

These data show that thymidylate synthase is among the most conserved proteins that have been examined. Because thymidylate synthase is essential for the biosynthesis of TMP, it was probably present in the earliest organisms having DNA genomes. The fact that the enzyme has the same basic structure in widely divergent organisms suggests that the structure of the enzyme was optimized very early during evolution and implies that many of the conserved amino acids play an important structural or functional role. The amino terminal region is probably less important for the maintenance of structure or function, since it is the least conserved region of the enzyme.

**Absence of 3' Untranslated Region on mRNA**

Surprisingly, the termination codon, UAA, was followed immediately by poly(A). The sequence of pMTS-4 was also determined in this region and was the same as that of pMTS-3 except that the poly(A) tail was 21 nucleotides long. If this is the real poly(A) tail of the mRNA, this is the first example of a eukaryotic nuclear-encoded

| Table 1 |
| Amino Acid Identity between Thymidylate Synthase Sequences of Mouse and Other Species |

<table>
<thead>
<tr>
<th>SPECIES</th>
<th>IDENTITY (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Overall</td>
</tr>
<tr>
<td>Human</td>
<td>90</td>
</tr>
<tr>
<td><em>Lactobacillus casei</em></td>
<td>50</td>
</tr>
<tr>
<td><em>E. coli</em></td>
<td>55</td>
</tr>
<tr>
<td>T4 phage</td>
<td>47</td>
</tr>
<tr>
<td>( \phi 3T ) of <em>Bacillus subtilis</em></td>
<td>40</td>
</tr>
</tbody>
</table>

**NOTE.**—Data are derived from fig. 3 and were determined by multiplying the no. of amino acid identities by 100 and dividing the product by the number of positions that are not gapped in either sequence.
mRNA that lacks a 3' noncoding region (except for the poly-A tail). Anderson et al. (1981, 1982) noted that in several human and bovine mitochondrial mRNAs, the UAA stop codon is followed immediately by the poly(A) tail. In these cases, the mitochondrial genes code for the T or TA of the TAA stop codon, and the addition of the poly(A) tail completes the TAA termination codon.

An alternative explanation for our observation is that the mouse thymidylate synthase mRNA may have a short oligo(A) region immediately after the UAA codon, followed by an untranslated region and the real poly(A) tail. During the cloning of thymidylate synthase cDNA, the oligo(dT) primer might have primed reverse transcription starting at the oligo(A) region rather than at the poly(A) tail. To address this possibility, it is necessary to analyze the sequence of the mouse thymidylate synthase gene in the vicinity of the poly(A) addition site. We have recently isolated the entire mouse thymidylate synthase gene from a BamHI genomic library (in preparation). Analysis of the 5' and 3' exons have confirmed the sequences shown in figure 2. Furthermore, an oligo(A) stretch was not found adjacent to the nonsense codon. Therefore it appears that the poly(A) region immediately adjacent to the UAA codon does represent the poly(A) tail of the mRNA.

The human thymidylate synthase mRNA has a 3' untranslated region of nearly 500 nucleotides (Takeishi et al. 1985). However, the human message is $\geq$1.6 kb in length (Takeishi et al. 1985) compared with $\sim$1.3 kb for the mouse message. This size difference may arise in part from the absence of a 3' untranslated region in the mouse message.

Although the mouse sequence does not have the AATAAA polyadenylation signal (Proudfoot and Brownlee 1976), there is a sequence ATTTAA 21 bases upstream of the oligo(A) on the cDNA, in the coding region of the message. This sequence is known to function as a signal for polyadenylation in several other mRNAs and is located at the proper position to serve as the signal (Berget 1984; Birnstiel et al. 1985). The sequences of the mouse and human thymidylate synthase mRNAs are identical from the ATTTAA up to the poly(A) stretch of the mouse mRNA and are extremely similar upstream from this region. Since the human mRNA is not polyadenylated at the same site as the mouse mRNA, this may indicate that sequences downstream of the poly(A) addition site are important for determining the site of cleavage and polyadenylation (Berget 1984; Birnstiel et al. 1985).

Since the content of thymidylate synthase mRNA in growth-stimulated mouse fibroblasts appears to be regulated in part by controlling the polyadenylation of this RNA species (Jehn et al. 1985b) and since poly(A) appears to be added immediately adjacent to the termination codon, analysis of the mechanism of poly(A) addition for this mRNA species is warranted.

Acknowledgments

We thank Drs. J. DeWille, C.-H. Jenh, and S. Pilistine for comments on the manuscript. This study was supported by Public Health Service grants GM29356, HL34129 (E.F.V.), and CA16058 and grant 83005 from the Ohio Cancer Research Associates. L.F.J. was supported by Faculty Research Award FRA210 from the American Cancer Society.

LITERATURE CITED


WALTER M. FITCH, reviewing editor

Received January 27, 1986.
Substitution rates in pseudogenes can be used to estimate the frequencies of different types of mutation on the assumption that pseudogenes are not subject to selective constraints. These rates are used here to investigate the effect of neighboring bases on mutation rates. There is a marked increase in the frequency of transitions, though not of transversions, from the doublet CG. There are also some smaller effects of neighboring bases on the frequencies of transitions from adenine and thymine. The results are used to predict dinucleotide frequencies in a stretch of DNA subject to no selective constraints and to investigate the possibility of non-randomness in the usage of stop codons.

Introduction

Since pseudogenes are apparently functionless (Li et al. 1981), substitution rates in them should be proportional to mutation rates—and can be used to estimate the relative frequencies of different types of mutation. Gojobori et al. (1982) and Li et al. (1984) have used this fact to estimate the pattern of point mutations in nucleotides and have observed a high frequency of transitions from CG dinucleotides, which can be explained by the observation that at this doublet in vertebrates there is a high level of methylated cytosine that mutates abnormally frequently to thymine (Coulondre et al. 1978; Bird 1980; Razin and Riggs 1980). The purpose here is to investigate systematically the effect of neighboring bases on substitution rates in pseudogenes in order to estimate the magnitude of the CG effect and to determine whether other neighbor effects exist.

Material and Methods


1. Key words: pseudogene, substitution rate, methylated cytosine, dinucleotide frequency, stop codon.

Address for correspondence and reprints: Dr. Michael Bulmer, Department of Biomathematics, University of Oxford, 5 South Parks Road, Oxford OX1 3UB, England.
Sequences were taken from the GenBank and EMBL nucleic acid data banks. Each pseudogene was aligned with the relevant functional gene, using the published alignment allowing for insertions and deletions as required and considering only the coding region of the gene. To ensure that the direction of any observed change was known, only those sites were used at which one could be reasonably confident that no change had occurred in the functional gene since the origin of the pseudogene. The following criteria were used to this effect: (1) All sites were used for the two RNA genes, following Gojobori et al. (1982), since nucleotide substitution is so slow in such genes that it is safe to assume that no substitutions have occurred in the functional genes. (2) For beta-tubulin and immunoglobulin C-epsilon only those sites in the functional gene were used at which a nucleotide substitution must cause an amino acid replacement, following Li et al. (1984). The amino acid sequence is highly conserved in these genes, so that it is safe to assume that no change has occurred in the functional gene at these sites, but it would be unsafe to make this assumption with respect to the remaining sites at which synonymous substitutions are possible. (3) For the more variable immunoglobulin V-kappa and globin genes, only those sites in the functional gene were included at which the observed nucleotide was the same as the consensus nucleotide for the sequences of two or three species (man and mouse for the immunoglobulin V-kappa gene; man, mouse, and rabbit for the alpha- and beta-globin genes); if no consensus (i.e., strict majority) existed, the site was excluded.

No obvious evidence of gene conversion was seen in any of the pseudogenes studied. Even if it had been present it would not have affected the conclusions reached about the relative frequencies of different substitutions.

Results

Table 1 shows substitutions summed over all 14 pseudogenes and over both the coding and the noncoding strands. (A preliminary analysis showed no difference between the two strands, as expected a priori. For example, transitions from A to G on the coding strand are about as frequent as those from T to C, which correspond with the transitions from A to G on the noncoding strand.) Only the first half of the table (for A and C) is shown, since the full table is point symmetric about the center. The two types of transversion (C or T from A, A or G from C) are about equally frequent and will be pooled in future discussion. Note that transitions are more frequent than transversions and that substitutions are more frequent from C or G than from A or T, in line with previous results (Gojobori et al. 1982; Li et al. 1984).

To assess the effect of a neighboring nucleotide, it is necessary to ensure that that nucleotide has not changed; in addition to the exclusions at the site itself, mentioned above, any sites that were followed by insertions or deletions in the pseudogene or for which the neighboring nucleotide differed in the functional gene and the pseudogene

Table 1
Base Substitutions in 14 Pseudogenes

<table>
<thead>
<tr>
<th>BASE IN FUNCTIONAL GENE</th>
<th>BASE IN PSEUODEogene</th>
<th>TRANSITIONS (%)</th>
<th>TRANSVERSIONS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2,216 47 106 36</td>
<td>4.4</td>
<td>3.5</td>
</tr>
<tr>
<td>C</td>
<td>79 2,274 69 283</td>
<td>10.5</td>
<td>5.5</td>
</tr>
</tbody>
</table>
were also excluded. The effect of the base on the right on substitutions in the pseudogene is shown in table 2. As before, data have been summed over all 14 pseudogenes and over both the coding and the noncoding strands. It is not necessary to consider separately the effect of the base on the left since this is given by symmetry; for example, the number of transitions from A with G on the right is equal to the number of transitions from T with C on the left.

There is no evidence of any effect of the neighboring base on the frequency of transversions. There is a large neighbor effect on transitions from C, one that results entirely from an increased frequency with G on the right; note that CG is self-complementary, so that there is a similar increase in transitions from G with C on the left. There are also smaller neighbor effects, which are similar in kind, on transitions from A or T; the transition frequency from either of these bases is reduced by having G on the right (or C on the left) and is increased by having T on the right (or A on the left).

The CG effect probably results from the occurrence of methylated cytosine in this doublet in vertebrates and accounts for the deficiency of CG doublets (Bird 1980). Since there is variability in the degree of CG doublet deficiency between different parts

<table>
<thead>
<tr>
<th>Base in Functional Gene</th>
<th>Base on Right</th>
<th>Transitions (%)</th>
<th>Transversions (%)</th>
<th>Total Bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A</td>
<td>4.0</td>
<td>2.5</td>
<td>399</td>
</tr>
<tr>
<td>A</td>
<td>C</td>
<td>4.6</td>
<td>1.8</td>
<td>565</td>
</tr>
<tr>
<td>A</td>
<td>G</td>
<td>2.4</td>
<td>2.1</td>
<td>672</td>
</tr>
<tr>
<td>A</td>
<td>T</td>
<td>6.1</td>
<td>4.3</td>
<td>396</td>
</tr>
</tbody>
</table>

\( \chi^2 \) for heterogeneity 9.5* 7.0 NS

<table>
<thead>
<tr>
<th>Base in Functional Gene</th>
<th>Base on Right</th>
<th>Transitions (%)</th>
<th>Transversions (%)</th>
<th>Total Bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>A</td>
<td>4.9</td>
<td>5.1</td>
<td>691</td>
</tr>
<tr>
<td>C</td>
<td>C</td>
<td>6.9</td>
<td>3.6</td>
<td>798</td>
</tr>
<tr>
<td>C</td>
<td>G</td>
<td>44.0</td>
<td>5.8</td>
<td>225</td>
</tr>
<tr>
<td>C</td>
<td>T</td>
<td>6.0</td>
<td>4.5</td>
<td>516</td>
</tr>
</tbody>
</table>

\( \chi^2 \) for heterogeneity 331.7** 2.8 NS

<table>
<thead>
<tr>
<th>Base in Functional Gene</th>
<th>Base on Right</th>
<th>Transitions (%)</th>
<th>Transversions (%)</th>
<th>Total Bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>G</td>
<td>A</td>
<td>10.4</td>
<td>6.3</td>
<td>511</td>
</tr>
<tr>
<td>G</td>
<td>C</td>
<td>9.1</td>
<td>4.7</td>
<td>616</td>
</tr>
<tr>
<td>G</td>
<td>G</td>
<td>10.1</td>
<td>3.8</td>
<td>685</td>
</tr>
<tr>
<td>G</td>
<td>T</td>
<td>10.8</td>
<td>5.0</td>
<td>517</td>
</tr>
</tbody>
</table>

\( \chi^2 \) for heterogeneity 1.0 NS 3.9 NS

<table>
<thead>
<tr>
<th>Base in Functional Gene</th>
<th>Base on Right</th>
<th>Transitions (%)</th>
<th>Transversions (%)</th>
<th>Total Bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>T</td>
<td>A</td>
<td>4.0</td>
<td>1.5</td>
<td>202</td>
</tr>
<tr>
<td>T</td>
<td>C</td>
<td>4.0</td>
<td>2.6</td>
<td>647</td>
</tr>
<tr>
<td>T</td>
<td>G</td>
<td>2.2</td>
<td>3.6</td>
<td>810</td>
</tr>
<tr>
<td>T</td>
<td>T</td>
<td>5.6</td>
<td>2.8</td>
<td>394</td>
</tr>
</tbody>
</table>

\( \chi^2 \) for heterogeneity 9.3* 2.9 NS

NOTE.—The \( \chi^2 \) values for heterogeneity test for an effect of the base on the right. NS = not significant.

* \( P < .05 \) on 3 degrees of freedom.

** \( P < .001 \) on 3 degrees of freedom.
of the genome (Smith et al. 1985), it is of interest to see whether there is any variability in the magnitude of the CG effect between pseudogenes derived from different functional genes. Table 3 shows that there is no evidence of heterogeneity, but the data are somewhat inconclusive since immunoglobulin V-kappa and beta-globin are the only genes in the table with a large CG-doublet deficiency.

For simplicity, the analyses in tables 1 and 2 have been performed on pooled data. Since it is possible under some circumstances to obtain spurious results when pooling contingency tables, a more complicated analysis was done on the unpooled data using a logistic regression model with the statistical package GLIM (McCullagh and Nelder 1983). The conclusions were identical. In particular, there was no difference between the behavior of processed and nonprocessed pseudogenes.

The results in table 2 can be used to estimate substitution rates (relative to an arbitrary time scale) as follows: (1) For transversions, there is no evidence of a neighbor effect, and the substitution rate can be estimated as the overall proportion of substitutions—which is 0.028 from A or T and 0.046 from C or G—divided equally between the two types of transversion. (2) For transitions from C, the rate can be taken as 0.060 with A, C, or T on the right, since there is no evidence of heterogeneity between these three bases on the right. The frequency of transitions from C with G on the right is 0.440, but this figure should be increased to 0.580 (= -ln[1−0.44]) in estimating the transition rate to allow for the nonlinearity in the graph of frequency of substitution plotted against time. (This calculation assumes that the proportion of unchanged CG doublets declines exponentially with time. The correction is unnecessary for the other, much smaller frequencies.) There is no effect of the base on the left on transitions from C, since there is no effect of the base on the right on transitions from G. By symmetry, the transition rate from G with A, G, or T on the left is 0.580; there is no effect of the base on the right. (3) For transitions from A or T there is a smaller effect of the base on both sides. (Remember that an effect of the base on the right of A implies an effect of the complementary base on the left of T, and vice versa.) The results in table 2 can be approximately reproduced by taking the transition rate from A or T with A or C on the right and G or T on the left as 0.042 and multiplying this rate by 0.55 with G on the right or C on the left or by 1.38 with T on the right or A on the left.

Discussion

The main conclusion is that there is a tenfold increase in the frequency of transitions in CG doublets in vertebrates, an increase that can be attributed to the high

Table 3
CG Effect for Different Genes

<table>
<thead>
<tr>
<th>GENE</th>
<th>NO. OF PSEUDOGENES</th>
<th>% TRANSITIONS FROM C (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>G on Right</td>
</tr>
<tr>
<td>Beta-tubulin</td>
<td>2</td>
<td>43.8 (73)</td>
</tr>
<tr>
<td>Xenopus 5S rRNA</td>
<td>1</td>
<td>33.3 (3)</td>
</tr>
<tr>
<td>U1 RNA</td>
<td>5</td>
<td>30.9 (58)</td>
</tr>
<tr>
<td>Immunoglobulin V-kappa</td>
<td>1</td>
<td>100.0 (2)</td>
</tr>
<tr>
<td>Immunoglobulin C-epsilon</td>
<td>1</td>
<td>46.4 (56)</td>
</tr>
<tr>
<td>Alpha-globin</td>
<td>2</td>
<td>60.6 (33)</td>
</tr>
<tr>
<td>Beta-globin</td>
<td>2</td>
<td>100.0 (2)</td>
</tr>
</tbody>
</table>
frequency of methylated cytosine in this doublet. There are also some smaller effects of neighboring bases on frequencies of transitions from A and T.

These neighbor effects will introduce nonrandomness into the sequence of nucleotides. Table 4 shows the predicted dinucleotide frequencies in a stretch of DNA with the substitution rates estimated above and subject to no selective constraints. It is AT rich, a characteristic observed in vertebrate noncoding regions (Gojobori et al. 1982). There is a reduction in CG doublets to about one fifth of the frequency expected under independence; there is an increase of \( \sim 25\% \) above expectation in the frequencies of CA, TG, AG, and CT doublets. These predictions agree qualitatively with observed dinucleotide frequencies in vertebrates (Bird 1980; Nussinov 1984), though no information is available for noncoding regions. The main exception is that the observed deficiency of TA doublets is not predicted in table 4. A model incorporating the CG effect but ignoring the neighbor effect for A and T predicts a similar decrease in CG doublets and a similar increase in CA and TG doublets (which are reached from CG by a single transition) but a smaller increase (by 9%) in AG and CT doublets. (The dinucleotide frequencies were condensed from the equilibrium distribution for trinucleotides, which was obtained by using the theory of continuous-time Markov processes [Cox and Miller 1965]. To obtain numerical results by an iterative procedure it was assumed that dependencies in the sequence do not extend farther than two bases either way; this is not exactly true but should provide a satisfactory approximation.)

These results are also of value in interpreting codon usage. Consider as an example the usage of the three stop codons TAG, TAA, and TGA, which will be numbered 1, 2, and 3. If \( m_{ij} \) is the rate of mutation from stop codon \( i \) to stop codon \( j \), the relative frequency, in the absence of selection, of stop codon \( i \), \( p_i \), satisfies the differential equation

\[
\frac{dp_i}{dt} = \sum_{j \neq i} m_{ji}p_j - \sum_{j \neq i} m_{ij}p_i.
\]

(1)

(Any mutation to a non-stop codon can be ignored, since it will be rapidly eliminated.) Setting these equations to zero and observing that \( m_{13} = m_{31} = 0 \), since TAG and TGA are not connected by a single step mutation, we find that the equilibrium frequencies should be

\[
\begin{align*}
p_1 &= m_{21} m_{32}/k, \\
p_2 &= m_{12} m_{32}/k, \\
p_3 &= m_{12} m_{23}/k, \\
( k &= m_{21} m_{32} + m_{12} m_{32} + m_{12} m_{23}).
\end{align*}
\]

(2)

From the results at the end of the last section, the mutation rates in vertebrates are estimated as

\[
\begin{align*}
m_{12} &= m_{32} = 0.06, \\
m_{21} &= 0.056, \\
m_{23} &= 0.040.
\end{align*}
\]

(3)

The equilibrium frequencies of the three stop codons in vertebrates should therefore be
Table 4
Predicted Dinucleotide Frequencies per Thousand

<table>
<thead>
<tr>
<th>BASE ON LEFT</th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>T</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>86</td>
<td>66</td>
<td>74</td>
<td>86</td>
<td>312</td>
</tr>
<tr>
<td>C</td>
<td>74</td>
<td>34</td>
<td>7</td>
<td>74</td>
<td>188</td>
</tr>
<tr>
<td>G</td>
<td>54</td>
<td>34</td>
<td>34</td>
<td>66</td>
<td>188</td>
</tr>
<tr>
<td>T</td>
<td>98</td>
<td>54</td>
<td>74</td>
<td>86</td>
<td>312</td>
</tr>
<tr>
<td>Total</td>
<td>312</td>
<td>188</td>
<td>188</td>
<td>312</td>
<td>1,000</td>
</tr>
</tbody>
</table>

\[ p_1 = 0.36, \]
\[ p_2 = 0.38, \]
\[ p_3 = 0.26. \]  

(4)

Table 5 shows the observed stop codon usage in a number of species. Only the Epstein-Barr virus seems to have the predicted random usage. All the other species avoid TAG, some having a preference for TGA, others for TAA. These facts suggest that selective pressures act on stop codon usage as they do on the usage of amino acid codons (Ikemura 1985). Two possible selective factors are (1) differential recognition by the release factors that catalyze termination and (2) differential misreading by tRNAs leading to readthrough. However, at least part of the preference for TGA in vertebrates may be due to the general tendency to avoid the doublet TA, a tendency that is not accounted for by mutation rates derived from table 2.

Table 5
Stop Codon Usage (%) in Different Species

<table>
<thead>
<tr>
<th>SPECIES (n)</th>
<th>TAG</th>
<th>TAA</th>
<th>TGA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Man (115)</td>
<td>17</td>
<td>29</td>
<td>55</td>
</tr>
<tr>
<td>Mouse (89)</td>
<td>19</td>
<td>26</td>
<td>55</td>
</tr>
<tr>
<td>Chicken (39)</td>
<td>15</td>
<td>44</td>
<td>41</td>
</tr>
<tr>
<td>Drosophila (27)</td>
<td>15</td>
<td>85</td>
<td>0</td>
</tr>
<tr>
<td>Yeast (40)</td>
<td>12</td>
<td>70</td>
<td>18</td>
</tr>
<tr>
<td>E. coli (154)</td>
<td>6</td>
<td>71</td>
<td>23</td>
</tr>
<tr>
<td>Phage lambda (60)</td>
<td>7</td>
<td>37</td>
<td>56</td>
</tr>
<tr>
<td>Phage T7 (56)</td>
<td>11</td>
<td>55</td>
<td>34</td>
</tr>
<tr>
<td>Epstein-Barr virus (58)</td>
<td>36</td>
<td>38</td>
<td>26</td>
</tr>
</tbody>
</table>

Acknowledgment

I thank Jasper Rees for help in accessing the GenBank and EMBL nucleic acid data banks.


MASATOSHI NEI, reviewing editor

Received November 7, 1985; revision received February 3, 1986.
Molecular Evolution of Mammalian Lactate Dehydrogenase-A Genes and Pseudogenes: Association of a Mouse Processed Pseudogene with a B1 Repetitive Sequence

Kayoko M. Fukasawa,* Wen-Hsiung Li,† Kiyohito Yagi,* Chi-Cheng Luo,† and Steven S.-L. Li*

*Laboratory of Genetics, National Institute of Environmental Health Sciences, National Institutes of Health; and †Center for Demographic and Population Genetics, University of Texas Health Science Center at Houston

A mouse genomic clone containing a lactate dehydrogenase-A (LDH-A) processed pseudogene and a B1 repetitive element was isolated, and a nucleotide sequence of ~3 kb was determined. The pseudogene and B1 element are flanked by perfect 13-bp repeats, and the B1 sequence starts at 14 nucleotides 3’ to the presumptive polyadenylation signal of the pseudogene. The nucleotide sequences of the LDH-A genes and processed pseudogenes from mouse, rat, and human were compared, and a phylogenetic tree was constructed. The rate and pattern of nucleotide substitutions in the LDH-A pseudogenes are similar to previously reported results (Li et al. 1984). The average rate of nucleotide substitutions in the LDH-A pseudogenes is $4.3 \times 10^{-9}$/site/year. The substitutions of C → T and G → A are most frequent, and A → G substitutions are relatively high. The rate of synonymous substitutions in the LDH-A genes is $5.3 \times 10^{-9}$, which is not significantly higher than the average rate of $4.7 \times 10^{-9}$ for 35 mammalian genes. The rate of nonsynonymous substitutions in the LDH-A genes is $0.20 \times 10^{-9}$, which is considerably lower than the average rate of $0.88 \times 10^{-9}$ for 35 mammalian genes. Thus, the mammalian LDH-A gene appears to be highly conserved in evolution.

Introduction

Lactate dehydrogenase (LDH; E.C. 1.1.27) catalyzes the interconversion of lactate and pyruvate with nicotinamide adenine nucleotide (NAD+) as coenzyme (Everse and Kaplan 1973). In mammals and birds, the five isozymes of tetrameric LDH are found in various proportions among different somatic tissues and are produced in vivo by combination of the A and B subunits, whereas the homotetrameric LDH-C4 is present only in mature testes and sperm (Market et al. 1975). The LDH-A4, LDH-B4, and LDH-C4 isozymes possess distinct physicochemical, catalytic, and immunological properties (Holbrook et al. 1975). The LDH-A (muscle), -B (heart), and -C (testis) polypeptides are encoded by three different gene loci that originated from an ancestral gene during the course of evolution (Market et al. 1975; Li et al. 1983). In order to know the structural and evolutionary relationships of the LDH-A, -B, and -C genes

1. Key words: LDH-A genes, pseudogenes, nucleotide substitution, rate of substitution, B1 repeat.

Address for correspondence and reprints: Dr. Steven S.-L. Li, Laboratory of Genetics, National Institute of Environmental Health Sciences, National Institutes of Health, Research Triangle Park, North Carolina 27709.

© 1986 by The University of Chicago. All rights reserved.
0737-4038/86/0304-3306$02.00
and to study the molecular mechanisms of gene regulation, we have undertaken the investigation of gene organization and protein structure of mammalian LDH isozymes. We have reported the nucleotide sequences of human LDH-A cDNA and a pseudogene (Tsujibo et al. 1985). We have also described the genomic organization of protein-coding exons of the LDH-A functional genes from mouse and humans (Chung et al. 1985; Li et al. 1985a). In this paper, we present the nucleotide sequence of a mouse LDH-A processed pseudogene associated with a B1 repetitive element complementary to the most abundant class of mouse fold-back RNA (Krayev et al. 1980), the rate of nucleotide substitution in mammalian LDH-A functional genes and processed pseudogenes, and the pattern of nucleotide substitutions in pseudogenes.

Material and Methods

Isolation and Sequencing of a Mouse LDH-A Pseudogene

The mouse Mus musculus C57/B10 LDH-A genomic clones were isolated from a Charon 4A genomic library provided by Dr. M. Edgell (Benton and Davis 1977; Rigby et al. 1977; Weaver et al. 1981). The DNA purified from the genomic clone λM10 was analyzed by restriction-endonuclease mapping and Southern blotting (Maniatis et al. 1982). The isolated DNA fragment containing the LDH-A processed pseudogene was further cleaved and subcloned into M13 mp10/mp11 phages (Messing et al. 1981). The nucleotide sequences of the genomic DNA fragments purified from M13 phages were determined by the dideoxy chain termination method (Sanger et al. 1977).

Sequence Data

The nucleotide sequence of the mouse processed pseudogene λM10 is determined in this study. The sequence data for the mouse LDH-A functional gene are from Akai et al. (1985) and Li et al. (1985a; unpublished data). The complete sequence of rat Ratus norvegicus LDH-A cDNA pLDH-2 was reported by Matrisian et al. (1985). The nucleotide sequences of human Homo sapiens LDH-A cDNA pCD380 and processed pseudogene λH463 were described previously (Tsujibo et al. 1985).

Statistical Methods

We considered only the amino acid-coding regions. The number of synonymous (causing no amino acid changes) and nonsynonymous (resulting in different amino acids) substitutions between each pair of genes was obtained by the method of Li et al. (1985b). The pattern of nucleotide substitutions in the human and mouse pseudogenes was obtained by following the procedure used in Gojobori et al. (1982) and Li et al. (1984). Since five sequences are available from humans and rodents, it is rather easy to deduce the ancestral sequence of a pseudogene and its functional counterpart and then to infer the nucleotide substitutions in the pseudogenes (Fitch 1971; Gojobori et al. 1982); in each case several nucleotide sites are excluded because the ancestral nucleotides cannot be determined uniquely or because deletions have occurred in the pseudogene studied. This exclusion should have little effect on the final results because in each comparison the number of sites excluded is less than four.

The origin of a processed pseudogene and its rate of nucleotide substitution are estimated by a method developed by W.-H. Li (unpublished data). The data are shown in figure 1. The processed pseudogene is denoted by \( \Psi \) and is assumed to have been nonfunctional since the time of its origin. Its functional counterpart is denoted by \( F \).
FIG. 1.—Model for estimating the time (t) of separation between a pseudogene (ψ) and its functional counterpart (F) and the rate (b) of nucleotide substitution in the pseudogene. R denotes the functional gene from a reference species; \( a_i \) denotes the rate of nucleotide substitution at the \( i \)th position of codons in the functional genes; and \( l_i, m_i, \) and \( n_i \) denote the number of nucleotide substitutions per site at the \( i \)th codon position between nodes O and ψ, O and F, and O and R, respectively.

and the functional gene from a reference species is denoted by R. The divergence time (T) between the two species is assumed to be known; \( T = 75 \) Myr for the divergence between humans and rodents. The rate of nucleotide substitution at the \( i \)th position of codons in the functional gene is then given by

\[
a_i = d_{FRi}/(2T),
\]  

where \( d_{FRi} \) is the number of substitutions per nucleotide site at the \( i \)th position of codons between genes F and R. In the figure \( l_i, m_i, \) and \( n_i \) denote the number of substitutions per site at the \( i \)th position of codons between O and ψ, between O and F, and between O and R, respectively. These branch lengths can be estimated by inferring the ancestral sequence at the node O or by the following equations:

\[
l_i = \frac{1}{2}[d_{\psi Fi} + d_{\psi Ri} - d_{FRi}],
\]

\[
m_i = \frac{1}{2}[d_{\psi Fi} - d_{\psi Ri} + d_{FRi}],
\]

\[
n_i = \frac{1}{2}[-d_{\psi Fi} + d_{\psi Ri} + d_{FRi}],
\]

(Li et al. 1981). Let \( A = \sum l_i \), \( B = \sum m_i \), and \( C = \sum n_i \), where the summations are over the three positions of codons, and let \( V(A) \), \( V(B) \), and \( V(C) \) be the variances of
$A$, $B$, and $C$, respectively. Then the mean and variance of the divergence time $t$ between $\psi$ and $F$ are approximately given by

$$t = \frac{2TB}{D},$$

$$V(t) = \frac{(2T)^2}{D^2} \left[ V(B) + \frac{B^2}{D^2} V(D) - \frac{2B}{D} V(B) \right],$$

where $D = B + C$, and $V(D) = V(B) + V(C)$. The mean and variance of the rate ($b$) of nucleotide substitution in the pseudogene are approximately given by

$$\bar{b} = \frac{AD}{6TB},$$

$$V(b) = \frac{1}{(6TB)^2} \left[ V(X) + \frac{X^2}{B^2} V(B) - \frac{XA}{B} V(B) \right],$$

where $X = (B + C)D$ and $V(X) = [V(B) + V(C)]V(D)$. The above variances were derived under the assumption that the three positions of codon evolved independently but that all the $i$th codon positions evolved at the same rate.

Results and Discussion

Nucleotide Sequence of a Mouse LDH-A Processed Pseudogene

Several mouse genomic clones containing LDH-A genelike sequences were isolated and partially characterized (Li et al. 1985a). A genomic clone, $\lambda$M10, which exhibited a strong hybridization signal to the LDH-A cDNA probes from mouse (Akai et al. 1985) and humans (Tsujibo et al. 1985), was shown to contain a 15-kb DNA insert. The restriction-endonuclease map of the genomic clone $\lambda$M10 along with the DNA sequencing strategy is shown in figure 2, and the nucleotide sequence of ~3 kb is given in figure 3. A sequence comparison of the genomic clone $\lambda$M10 sequence with the mouse LDH-A functional gene (Li et al. 1985a) indicated that the genomic clone $\lambda$M10 contains a processed LDH-A pseudogene and a B1 repetitive element. As indicated in figure 4, the nucleotide sequence of the mouse LDH-A pseudogene $\lambda$M10 possesses all the characteristics of a processed pseudogene—i.e., (1) absence of all seven introns present in LDH-A functional genes, (2) nucleotide substitutions at codons 54 and 65 resulting in chain termination codons, and (3) two single-base deletions in codons 135 and 215 and a deletion of four nucleotides in codons 10 and 11. Although the nucleotide sequence of this pseudogene ($\lambda$M10) cannot code for a functional protein, it has 92% identity with the protein-coding sequence of the functional mouse LDH-A gene ($\lambda$M15).

The mouse genomic clone $\lambda$M10 contains two perfect 13-bp repeats, one located just upstream from the 5' untranslated region of the LDH-A processed pseudogene and the other downstream from the A-rich sequence of the B1 repetitive element (fig. 3). The nucleotide sequence upstream from the 5' direct repeat of this pseudogene is very different from that of the promoter region of mouse LDH-A functional gene $\lambda$M15 (S. S.-L. Li, unpublished data). Thus, this LDH-A processed pseudogene is transcriptionally inactive, since it lacks the necessary 5' regulatory sequences. It has been proposed that intronless pseudogenes were derived from the processed mRNA intermediates through retrovirus-like transposable elements, since these pseudogenes
are often flanked by direct repeated DNA sequences (Lueders et al. 1982). In higher eukaryotes, such mechanisms of reverse copying and the presence of many processed pseudogenes have been described (Proudfoot 1980; Little 1982; Sharp 1983).

The B1 repetitive element present in mouse LDH-A pseudogene λM10 starts 14 nucleotides 3' to the presumptive polyadenylation signal (fig. 3) (Proudfoot and Brownlee 1976), and only eight of 132 nucleotides compared are different from the consensus sequence of mouse B1 repetitive elements, as indicated in figure 5 (Kalb et al. 1983). The structures homologous to the RNA polymerase III control region, the Hogness-Goldberg box, and the major T antigen-binding site are also conserved in the B1 element present in mouse LDH-A pseudogene λM10. The mouse B1 repetitive sequences of ~130 nucleotides are homologous to the human Alu families formed from two similar direct repeats each ~130 bases long (Krayev et al. 1980, 1982; Deininger et al. 1981; Schmid and Jelinek 1982; Kalb et al. 1983).

Evolution of LDII-A Genes and Pseudogenes

Table 1 shows the number of substitutions per synonymous site ($K_S$) and the number of substitutions per nonsynonymous site ($K_A$) between each pair of sequences. The $K_S$ value between mouse λM15 and human pCD380 is identical to that between rat pLDH-2 and human pCD380. From this $K_S$ value we estimate that the rate of synonymous substitution in the LDH-A gene is $(0.80 \pm 0.10)/(2 \times 75 \times 10^6) = (5.3 \pm 0.7) \times 10^{-9}/\text{synonymous site/year}$. This rate is slightly, but not significantly, higher than the average synonymous rate $(4.7 \times 10^{-9})$ for 35 mammalian genes (Li et al. 1985b). The $K_A$ value between mouse λM15 and human pCD380 is also identical with that between rat pLDH-2 and human pCD380. The nonsynonymous rate estimated from this value is $(0.03 \pm 0.01)/(2 \times 75 \times 10^6) = (0.20 \pm 0.07) \times 10^{-9}/\text{nonsynonymous site/year}$.
FIG. 3.—Nucleotide sequences of the mouse LDH-A pseudogene λ10, a B1 repetitive element, and their flanking regions. The nucleotide sequence of the mouse LDH-A pseudogene λ10 is numbered from the presumptive initiation site of the "coding region," and the number of the last nucleotide in each row is given at the right-hand side. The start sites for 5' and 3' untranslated regions and for the B1 repetitive element are denoted by WT, 3'UT, and B1, respectively. The presumptive polyadenylation signal of AATGAA is also indicated. The flanking sequences of the LDH-A pseudogene and B1 repetitive element are given in lowercase, and the perfect 13-bp repeats are denoted by arrows. The solid-black arrowheads indicate the positions of the last six of the seven introns in the mouse LDH-A functional gene. The position of the first intron in the 5' untranslated region is uncertain.

This is considerably lower than the average nonsynonymous rate (0.88 \times 10^{-3}) for 35 mammalian genes (Li et al. 1985). Thus, the structure of the mammalian LDH-A gene appears to have been highly conserved in evolution.

It is clear from table 1 that the human pseudogene hH463 arose after the diver-
Fig. 4.—Comparison of nucleotide and amino acid sequences of the coding region of the mouse LDH-A gene λM15 and the processed pseudogene λM10. The nucleotide sequence of the mouse LDH-A pseudogene (λM10) is compared with the protein-coding region of the mouse LDH-A functional gene (λM15). Different nucleotides are given; identical nucleotides are indicated by dots, and deletions are denoted by dashes. The deduced amino acid sequence of mouse LDH-A gene is given above its nucleotide sequence. Only the different amino acids, terminations (Ter), and deletions (Del) deduced from the pseudogene AM10 are indicated.

gence between humans and rodents, because both the $K_S$ and $K_A$ values between human $\lambda H463$ and human pCD380 are smaller than the corresponding values between human $\lambda H463$ and either mouse $\lambda M15$ or rat pLDH-2. We also note that both the $K_S$ and $K_A$ values between the mouse pseudogene $\lambda M10$ and mouse $\lambda M15$ are smaller than the corresponding values between mouse $\lambda M10$ and rat pLDH-2. Therefore, mouse $\lambda M10$ probably arose after the mouse-rat divergence. From the preceding arguments, we propose the phylogenetic tree shown in figure 6. We assume that the root of the tree is the same as that in the human-rodent split and that the rate of nucleotide substitution is constant along each sequence linkage.

The model shown in figure 1 was used to estimate the time ($t$) since the divergence between the mouse pseudogene $\lambda M10$ and mouse $\lambda M15$ with the human functional gene as the R gene. We used all five sequences to infer the ancestral sequences of mouse $\lambda M15$ and $\lambda M10$ and then inferred the branch lengths $z_i$, $m_i$, and $n_i$; when a substitution could not be unambiguously assigned to one of the two branches involved,
it was split equally between the two branches. The inferred $I_i$, $m_i$, and $n_i$ values were then corrected for multiple hits by the method of Jukes and Cantor (1969) and Kimura and Ohta (1972). The inferred branch lengths $l_i$, $m_i$, and $n_i$ are shown in table 2. Entering these values of $A = \sum l_i$, $B = \sum m_i$, $C = \sum n_i$, $D = B + C$, and $T = 75$ Myr into equations (3)–(6), we obtain a $t$ of 18 ± 4 Myr between mouse $\lambda M10$ and $\lambda M15$ and a substitution rate ($b$) of $(3.6 \pm 0.7) \times 10^{-9}$ in the pseudogene $\lambda M10$ (table 3). This rate is substantially lower than the average rate $(4.7 \times 10^{-9})$ for mammalian pseudogenes (Li 1983).

Using the same procedure, we have also estimated the $t$ between the human pseudogene $\lambda H463$ and functional gene pCD380 and the $b$ in the pseudogene. In this case, however, we have used both mouse $\lambda M10$ and rat pLDH-2 as references. The inferred $I_i$, $m_i$, and $n_i$ values are shown in table 2, and the $t$ and $b$ values are given in table 3. The $t$ is 16 ± 4 Myr, and the substitution rate is $(4.9 \pm 1.0) \times 10^{-9}$. This rate is close to the average rate for mammalian pseudogenes.

**Table 1**

<table>
<thead>
<tr>
<th>Number of Synonymous Substitutions per Synonymous Site (above diagonal) and Number of Nonsynonymous Substitutions per Nonsynonymous Site (below diagonal) between Each Pair of Sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse $\lambda M15$</td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Mouse $\lambda M15$</td>
</tr>
<tr>
<td>Mouse $\lambda M10$</td>
</tr>
<tr>
<td>Rat pLDH-2</td>
</tr>
<tr>
<td>Human pCD380</td>
</tr>
<tr>
<td>Human $\lambda H463$</td>
</tr>
</tbody>
</table>

**NOTE.**—Mouse $\lambda M15$, rat pLDH-2, and human pCD380 are functional genes, and mouse $\lambda M10$ and human $\lambda H463$ are pseudogenes. The average number of synonymous and nonsynonymous sites in the functional genes is $\sim 217$ and $\sim 776$, respectively.
### Table 2
Estimates of the Branch Lengths in Figure 1

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$l_1$</th>
<th>$l_2$</th>
<th>$l_3$</th>
<th>$m_1$</th>
<th>$m_2$</th>
<th>$m_3$</th>
<th>$n_1$</th>
<th>$n_2$</th>
<th>$n_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1: Pseudogene = Mouse λM10; Gene $F$ = Mouse λM15; and Gene $R$ = Human pCD380</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d$</td>
<td>15</td>
<td>24</td>
<td>22</td>
<td>1</td>
<td>0</td>
<td>19</td>
<td>20</td>
<td>4</td>
<td>101</td>
</tr>
<tr>
<td>$n$</td>
<td>330</td>
<td>329</td>
<td>329</td>
<td>330</td>
<td>329</td>
<td>329</td>
<td>330</td>
<td>329</td>
<td>329</td>
</tr>
<tr>
<td>$s$</td>
<td>$4.7 \pm 1.2$</td>
<td>$7.7 \pm 1.5$</td>
<td>$7.0 \pm 1.5$</td>
<td>$0.3 \pm 0.3$</td>
<td>$0 \pm 0$</td>
<td>$6.0 \pm 1.4$</td>
<td>$6.3 \pm 1.4$</td>
<td>$1.2 \pm 0.6$</td>
<td>$39.5 \pm 4.3$</td>
</tr>
<tr>
<td>Case 2: Pseudogene = Human λH463; Gene $F$ = Human pDC380; and Gene $R$ = Mouse λM15 and Rat pLDH-2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$d$</td>
<td>26</td>
<td>25</td>
<td>24.5</td>
<td>4</td>
<td>1</td>
<td>13.5</td>
<td>17.5</td>
<td>3.5</td>
<td>103</td>
</tr>
<tr>
<td>$n$</td>
<td>329</td>
<td>329</td>
<td>329</td>
<td>329</td>
<td>329</td>
<td>329</td>
<td>329</td>
<td>329</td>
<td>329</td>
</tr>
<tr>
<td>$s$</td>
<td>$8.4 \pm 1.7$</td>
<td>$8.0 \pm 1.6$</td>
<td>$7.8 \pm 1.6$</td>
<td>$1.2 \pm 0.6$</td>
<td>$0.3 \pm 0.3$</td>
<td>$4.2 \pm 1.2$</td>
<td>$5.5 \pm 1.3$</td>
<td>$1.1 \pm 0.6$</td>
<td>$40.5 \pm 4.4$</td>
</tr>
</tbody>
</table>

**Note:** $d$ = The number of nucleotide differences; $n$ = the number of nucleotide sites; $s$ = the number of nucleotide substitutions per 100 sites. $s = -(300/4)\log[1 - (4d/n)]$ according to the method of Jukes and Cantor (1969) and Kimura and Ohta (1972).
FIG. 6.—A phylogenetic tree for the human LDH-A gene (pCD380) and pseudogene λH463, the rat LDH-A gene (pLDH-2), and the mouse LDH-A gene (λM15) and pseudogene λM10. The divergence time between mouse and rat is unknown and has been tentatively assumed to be 20 Myr.

In the model in figure 1, we assume that a processed pseudogene is nonfunctional from the time of its origin. Under this assumption, we should have $I_1 = I_2 = I_3$ because none of the three codon positions is subject to functional constraint and all would therefore evolve at the same rate. In the case of the human pseudogene λH463, the $I_1$, $I_2$, and $I_3$ are very similar (table 2). In the case of the mouse pseudogene, $I_2$ and $I_3$ are similar, although $I_1$ is considerably lower; the low $I_1$ value could represent an extreme random deviate.

Pattern of Nucleotide Substitutions in Pseudogenes

The pattern of nucleotide substitutions inferred from the mouse pseudogene λM10 is shown in the first matrix in table 4, while the pattern inferred from the human

Table 3
Time in Millions of Years since Separation between a Pseudogene and the Functional Gene ($t$), and Rates of Nucleotide Substitution for Mouse λM10 and Human λH463 Pseudogenes

<table>
<thead>
<tr>
<th>Pseudogene</th>
<th>$t$ (Myr)</th>
<th>$b$ ($\times 10^{-9}$)</th>
<th>$a_1$ ($\times 10^{-9}$)</th>
<th>$a_2$ ($\times 10^{-9}$)</th>
<th>$a_3$ ($\times 10^{-9}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse λM10</td>
<td>18 ± 4</td>
<td>3.6 ± 0.7</td>
<td>0.44 ± 0.10</td>
<td>0.08 ± 0.04</td>
<td>2.98 ± 0.30</td>
</tr>
<tr>
<td>Human λH463</td>
<td>16 ± 4</td>
<td>4.9 ± 1.0</td>
<td>0.45 ± 0.10</td>
<td>0.09 ± 0.04</td>
<td>3.03 ± 0.30</td>
</tr>
</tbody>
</table>

Note.—$a_1$, $a_2$, and $a_3$ are, respectively, the rate of nucleotide substitution per site per year at the first, second, and third positions of codons in the functional gene. $b$ is the value for all sites collectively.
### Table 4
Relative Substitution Frequencies (%) in Pseudogenes

<table>
<thead>
<tr>
<th></th>
<th>MOUSE λM10</th>
<th></th>
<th></th>
<th>HUMAN λH463</th>
<th></th>
<th></th>
<th>AVERAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>T</td>
<td>C</td>
<td>G</td>
<td>A</td>
<td>T</td>
<td>C</td>
</tr>
<tr>
<td>A</td>
<td>0.02 (1/251)</td>
<td>0.07 (4/251)</td>
<td>0.07 (4/251)</td>
<td>0.09 (7/266)</td>
<td>0.00 (0/266)</td>
<td>0.15 (12/266)</td>
<td>0.05</td>
</tr>
<tr>
<td>T</td>
<td>0.06 (3/227)</td>
<td>0.09 (5/277)</td>
<td>0.11 (6/227)</td>
<td>0.03 (2/263)</td>
<td>0.03 (2/263)</td>
<td>0.01 (1/263)</td>
<td>0.04</td>
</tr>
<tr>
<td>C</td>
<td>0.08 (4/225)</td>
<td>0.17 (9/255)</td>
<td>0.08 (4/225)</td>
<td>0.05 (3/191)</td>
<td>0.31 (18/191)</td>
<td>0.03 (2/191)</td>
<td>0.06</td>
</tr>
<tr>
<td>G</td>
<td>0.17 (11/285)</td>
<td>0.03 (2/285)</td>
<td>0.06 (4/285)</td>
<td>0.18 (14/262)</td>
<td>0.08 (6/262)</td>
<td>0.05 (4/262)</td>
<td>0.17</td>
</tr>
</tbody>
</table>

**NOTE.**—The substitutions are from the nucleotides in the left column to the nucleotides in the row. The values in parenthesis are the observed number of changes divided by the number of the nucleotides of the original type in the ancestral sequence. The relative substitution frequencies were calculated by the procedure of Gojobori et al. (1982).
pseudogene λH463 is shown in the second matrix. The two patterns are not very consistent, but this may be due to the small numbers (57 and 71, respectively) of nucleotide substitutions used. The third matrix is a simple average of the two patterns. The relative substitution frequencies in the third matrix are not very different from those inferred from 13 pseudogenes (Li et al. 1984). For example, C → T and G → A are the most frequent and A → G is relatively high. The sum of the relative frequencies of the four transitions is 0.24 + 0.17 + 0.11 + 0.06 = 0.58, which is close to the average (0.59) for 13 pseudogenes (Li et al. 1984).

Acknowledgments

We thank Mr. H. F. Tiano for technical assistance, Dr. M. Edgell for the mouse genomic library, Dr. B. H. Judd for his interest and support of this project, Dr. M. Komaki for her help in statistical analyses, and Drs. F. Johnson and C. Langley for reading the manuscript.

LITERATURE CITED


KRAYEV, A. S., K. G. SKRYABIN, A. P. BAYEV, and G. P. GEORGIJEV. 1980. The nucleotide sequence of the ubiquitous repetitive DNA sequence B1 complementary to the most abundant class of mouse fold-back RNA. Nucleic Acids Res. 8:1201–1215.


Walter M. Fitch, reviewing editor

Received October 28, 1985; revision received January 16, 1986.
The Simple Repeat Poly(dT-dG) • Poly(dC-dA) Common to Eukaryotes Is Absent from Eubacteria and Archaeabacteria and Rare in Protozoans

Julie Morris, Sidney R. Kushner, and Robert Ivarie
Department of Genetics, University of Georgia

Genomic DNA from a wide variety of prokaryotic and eukaryotic organisms has been assayed for the simple repeat sequence poly(dT-dG) • poly(dC-dA) by Southern blotting and DNA slot blot hybridizations. Consistent with findings of others, we have found the simple alternating sequence to be present in multiple copies in all organisms in the animal kingdom (e.g., mammals, reptiles, amphibians, fish, crustaceans, insects, jellyfish, nematodes). The TG element was also found in lower eukaryotes (Saccharomyces cerevisiae, Neurospora crassa, and Dictyostelium discoideum) and at a much lower frequency in protozoans (Oxytricha fallax and Tetrahymena thermophila). The sequence was also repeated in high copy number in a higher plant (Zea mays) as well as at very high levels in a unicellular green alga (Chlamydomonas reinhardtii). Although the copy number of the repeat per haploid genome was generally proportional to genome size, there was a greater-than-1,000-fold variation in the number of (TG)_{25}/100-kb genomic DNA. By contrast, no eubacterium—including Myxococcus xanthus, whose life cycle is very similar to that of the slime mold Dictyostelium discoideum, and Halobacter volcanii, whose genome contains other repeated sequences—was found whose genomic DNA contained this sequence in detectable amounts. A computer search also failed to find the TG element in human mitochondrial DNA.

Introduction

Poly(dT-dG) • poly(dC-dA) is a Z DNA-forming sequence that has recently been found in multiple copies in the eukaryotic genome (Nishioka and Leder 1980; Slightom et al. 1980; Miesfeld et al. 1981; Shen et al. 1981; Hamada and Kakunaga 1982; Hamada et al. 1982; Rogers 1983; Sun et al. 1984). The first TG elements of substantial length were discovered in the 3' flanking region of a mouse kappa gene (Nishioka and Leder 1980) and in the nontranscribed spacer region of the mouse ribosomal RNA genes (Miesfeld et al. 1981). The element has since been shown to be present in the genomes of widely separated eukaryotes from mammals to yeast (Miesfeld et al. 1981; Hamada and Kakunaga 1982; Hamada et al. 1982). In its most common form, thymine and guanine residues alternate almost perfectly on one strand of the helix for 30–60 bp. In humans, the element is present at 50,000 copies/haploid genome interspersed throughout the genomic DNA (Hamada and Kakunaga 1982; Hamada et al. 1982; Sun et al. 1984). The element does not contain terminal repeats (Hamada et al. 1984a; Sun et al. 1984) and can reside in 5' or 3' flanking regions and in untranslated regions of genes but not in coding sequences (Nishioka and Leder 1980; Slightom et al. 1980; Sun et al. 1984).

1. Key words: repeat sequence, poly(dT-dG) • poly(dC-dA), eubacteria, archaeabacteria, eukaryotes, evolution.

Address for correspondence and reprints: Dr. Robert Ivarie, Department of Genetics, University of Georgia, Athens, Georgia 30602.

© 1986 by The University of Chicago. All rights reserved.
0737-4038/86/0304-3404$02.00

Its presence in "lower" eukaryotes such as yeast (Miesfeld et al. 1981; Hamada and Kakunaga 1982; Hamada et al. 1982) and slime mold (Miesfeld et al. 1981) as well as in "higher" eukaryotes suggests that it may serve a role(s) in the structure and function of the eukaryotic genome. Although its role is not understood, TG elements have been proposed to be a hot spot for recombination and gene conversion (Slightom et al. 1980; Shen et al. 1981); they undergo an enhanced level of homologous recombination in oversized SV40 DNA in CVI cells relative to comparable SV40 homologous repeats (Stringer 1985). The repeat element can also act as a transcriptional enhancer in a transient expression vector (Hamada et al. 1984b) for genes transcribed by RNA polymerase II. However, RNA polymerase III transcription was inhibited by the presence of a long element in regions flanking a tRNA\textsuperscript{pro} gene from \textit{Caenorhabditis elegans} (Santoro et al. 1984) and by a 9-bp TG,CG sequence beginning 20 bp upstream from a \textit{Xenopus laevis} tRNA\textsuperscript{met} gene (Hipskind and Clarkson 1983).

The observations of Hamada et al. (1982) suggested that the number of TG repeats in a particular organism was proportional to its haploid genome size. However, only a few organisms were analyzed and none were prokaryotes. In this study we have examined a large number of eukaryotic and prokaryotic organisms for the sequence and iteration frequency.

Material and Methods

Genomic DNA


Genomic DNA was extracted from other organisms as follows: Nuclei from axenic amoebae of \textit{Dictyostelium discoideum} were isolated as described (Sogin and Olsen 1980), except that 0.2% Nonidet P-40 (Particle Data Laboratories, Ltd.) was used to lyse the cells; HeLa cell nuclei were isolated by low-speed centrifugation after lysis in 50 mM Tris(hydroxymethyl)aminomethane hydrochloride (Tris-HCl, pH 7.8), 0.1 M NaCl, 5 mM MgCl\textsubscript{2}, and 0.5% Nonidet P-40. Isolated nuclei or harvested sperm from \textit{Arbacia punctulata} (Gulf Specimens, Inc.) were then lysed by the addition of sodium dodecyl sulfate (SDS) to 0.4% and digested with 20 µg/ml protease K at 55°C for 4–8 h. \textit{Escherichia coli} were lysed by lysozyme–ethylenediamine tetraacetic acid treatment (Maniatis et al. 1982, p. 88), while \textit{Halobacter volcanii} were lysed osmotically by gentle resuspension in 25 mM Tris-HCl (pH 7.8), 0.1 M NaCl, and 5 mM ethylenediamine tetraacetic acid. DNA was isolated from all lysates by extracting twice with redistilled phenol, twice with chloroform:isoamyl alcohol (24:1), and once with chloroform. DNA was precipitated with 70% ethanol and collected by low-speed cen-
trifugation or spooling. After dissolving in TE (10 mM Tris-HCl [pH 8.0] and 1 mM ethylenediamine tetraacetic acid), the DNA was digested with heat-treated RNase (20 µg/ml) for 60 min at 37 °C and then extracted once each with phenol and chloroform: isoamyl alcohol (24:1).

Southern Blots

DNAs were digested with various restriction enzymes, electrophoresed on 0.8% agarose gels in 40 mM Tris-acetate (pH 8.0), and transferred to nitrocellulose filters (Schleicher and Schuell) as described (Ivarie et al. 1983). One microgram of each DNA sample was electrophoresed in each lane. Filters were prehybridized for 16 h at 43 °C in 5 X SSPE (1 X SSPE = 0.18 M NaCl, 10 mM NaH₂PO₄ • H₂O, 0.008% NaOH, and 1 mM ethylenediamine tetraacetic acid), 5 X Denhardt's solution (1 mg/ml each polyvinylpyrrolidone [Sigma], Ficoll 400 [Sigma], and bovine serum albumin), 50% formamide, and 100 µg/ml yeast tRNA and then hybridized to [³⁵S]-labeled poly(dT-dG) • poly(dC-dA) in the above solution for 24 h at 43 °C. The labeled probe was prepared by nick translation using [³⁵S]-(alpha-thio)-dATP as described (Ivarie et al. 1982; specific activity averaged 5 X 10⁸ cpm/µg). Blots were rinsed three times in 2 X SSPE and 0.1% SDS at 65 °C. To assay DNA for TG elements at lower stringency, hybridization was performed in 40% formamide at 40 °C and the last rinse of the filter was 2 X SSPE and 0.1% SDS at room temperature. Blots were visualized by autoradiography using Kodak XAR-5 X-ray film.

DNA Slot-Blot Hybridization

DNAs were denatured in 0.3 M NaOH for 60 min at 65 °C, neutralized with an equal volume of 2 M ammonium acetate, and then spotted onto nitrocellulose filters using a Schleicher and Schuell minifold II slot bloter. Filters were probed as described above for Southern blots at high or low stringency as noted in the legend to each figure. Autoradiographic intensities of each DNA slot were measured at an optical density of 500 nm in a Beckman DU-8 spectrophotometer and compared to a known standard as described in Results. Blots were exposed to X-ray film for varying times to ensure that optical densities fell within the linear-response range of the film.

Results

Southern Blots

Genomic DNA from a variety of organisms was digested with EcoRI or PstI and assayed for TG elements by Southern blotting and hybridizing to [³⁵S]-labeled poly(dT-dG) • poly(dC-dA). Typical results are illustrated in figure 1. Consistent with the findings of others (Miesfeld et al. 1981; Hamada and Kakunaga 1982; Hamada et al. 1982; Sun et al. 1984), human tumor cell DNA (HeLa) contained large amounts of the repeat sequence whereas Saccharomyces cerevisiae DNA contained fewer elements found in discrete restriction-enzyme bands. TG elements were also found in another fungus, Neurospora crassa, and at very high levels in the ciliated alga Chlamydomonas reinhardi. The element was also found in Zea mays (a flowering plant), Aurelia victoria (a jellyfish), and Drosophila melanogaster. Thus, by these limited criteria, the element appears to be ubiquitous in eukaryotic organisms.

By contrast, no detectable signal could be seen in the digests of total genomic DNA from the ciliated protozoan Tetrahymena thermophila or from three eubacteria (Escherichia coli, Bacillus subtilis and Myxococcus xanthus), two archaeabacteria.
(Methanococcus maripaludis and Halobacter volcanii), and the cyanobacterium Ana-
cystis nidulans (data not shown). The absence of signal from these organisms could reflect either the absence of TG elements from their genomes or sequence degeneracy such that they were not detected at the moderately high stringency conditions of hybridization. In the case of Tetrahymena, the element might also be confined to the micronucleus; >95% of total Tetrahymena DNA comes from the macronucleus, which has lost 10%–15% of the sequence complexity of micronuclear DNA (Yao and Goro

Southern Blots at Lower Stringency

To test these possibilities, micronuclear, macronuclear, and total Tetrahymena DNA as well as bacterial DNAs were digested with HindIII and assayed for TG elements
by Southern blotting and hybridization at lower stringency (see Material and Methods). These conditions are routinely used to assay single-copy sequences in *Tetrahymena* DNA, given its low (25%) G + C content. The results are illustrated in figure 2. Discrete bands were seen in total *Tetrahymena* DNA after digestion with HindIII, while DNA from both the macro- and micronucleus gave positive signals even though neither was digested to completion. It also appears that micronuclear DNA contained substantially more TG elements than did macronuclear DNA.

It can also be seen in figure 2 that prokaryotic DNAs were largely devoid of detectable TG sequences; faint smears were visible in undigested DNA regions of the gel for all bacterial DNAs except *E. coli* DNA, which gave no signal. In this regard, bacteriophage lambda also gave weak signals under these conditions of hybridization (data not shown), and a computer-assisted sequence search found that the largest TG element in lambda DNA is (TG)_4, which occurs three times. In more than 150,000 nucleotides of *E. coli* DNA, only a small (TG)_4 was found by computer search. A similar result was found by scanning 11,730 bp of cyanobacterial DNA, while 11,700 bp of halobacterial DNA contained two short sequences, (CA)_4 and TTCA and

![Southern blot hybridization of *Tetrahymena* and bacterial DNAs at lower stringency. Genomic DNA was digested with HindIII and assayed for TG elements by Southern hybridization as in fig. 1 but at lower stringency (see Methods). 1, *Tetrahymena* (total); 2, *Tetrahymena* micronucleus; 3, *Tetrahymena* macronucleus; 4, *Escherichia coli*; 5, *Methanococcus maripaludis*; 6, *Halobacter volcanii*; 7, *Pseudomonas putida*; 8, *Dictyostelium discoideum.*]
CG(CA)$_3$CGCA, that might have been able to give a hybridization signal at low stringency. Although the exact size and sequence degeneracy that can give a positive signal under low stringency conditions is not known, a short CA-rich sequence in the rat prolactin gene consisting of (CA)$_n$GA[(CA)$_2$GA]$_4$ in intron C (Cooke and Baxter 1982; Maurer 1985) gave a positive signal comparable to that for bacterial DNAs under the same conditions of hybridization (J. Morris and I. Farrance, unpublished data). These observations suggest that prokaryotic DNA may contain very short TG runs that are assayable at low stringency.

Repetition Frequency of TG Elements

The conclusion that the copy number of the conserved TG element in eukaryotes was proportional to genome size was based on a small sample of eukaryotic DNAs (Hamada and Kakunaga 1982; Hamada et al. 1982). Accordingly, we have tested the general validity of the conclusion by assaying for the presence of the TG elements in 20 eukaryotes and measuring the copy number for 17 of those for which genome sizes were readily available (Yao et al. 1974; Fasman 1976; Lewin 1980). One test of the hypothesis involved small bacterial genomes, and, as already shown above, such organisms lacked detectable elements. The eukaryotes assayed included simple ones such as fungi and protozoans as well as complex metazoans in the plant and animal kingdoms. Genome sizes ranged from $10^7$ bp in Saccharomyces to $5 \times 10^9$ bp in the southern toad, Bufo terrestris.

Reiteration frequency was estimated by quantitative analysis of DNA slot blots of serial dilutions of the various DNAs (figs. 3, 4) and by comparison with a standard containing a known level of TG elements. A lambda-genomic clone of the rat prolactin gene (Cooke and Baxter 1982) was used as a standard; it contains two TG elements in its 5' flanking region (Maurer 1985). Allowing for 10% sequence mismatching, the recombinant phage contained 170 bp of assayable TG elements in 51,500 bp.

An example of a serial dilution of the standard is shown in figure 4 (col. 1). Slot blots of the standard DNA gave a linear optical-density curve proportional to time of exposure to X-ray film and to the amount of DNA blotted, as long as optical density did not exceed 0.35. Hence, optical-density values less than 0.35 were used in calculating the TG-element levels in the DNAs. Copy number was based on an average TG length of 50 bp in accordance with Hamada et al. (1982). As table 1 shows, copy numbers agreed well with published values (Hamada and Kakunaga 1982; Hamada et al. 1982): 48,000 versus 50,000 for human and HeLa cell DNA; 2 $\times$ 10$^5$ versus 10$^5$ for mouse DNA; 500 versus 2,000 for Drosophila DNA; and 40 versus 100 for fungal DNA. Sun et al. (1984) estimated the frequency of TG elements as being 10$^9$/haploid genome assuming a 40-bp length. It appears, therefore, that the rat prolactin gene provided a suitable standard for estimating genomic frequencies of this repeat family.

The TG-element frequency is plotted against genome size in figure 5. In general, reiteration frequency increased as the genome size increased. However, there was wide variation in the number of (TG)$_{35}$/100 kb of genomic DNA (table 1), ranging from 0.005 in protozoans to 25 in Chlamydomonas with a mean ± SE value of 3.9 ± 1.4. Even when the outlying values are ignored, the range was quite wide, being 0.1–13.2. Species with large genomes and low ratios included the catfish Bagre marinus, the toad Bufo terrestris, and the flowering plant Zea mays. Furthermore, “lower” eukaryotes (e.g., fungi and protozoans) had ratios <0.15, yet the single-celled ciliate Chlamydomonas had the highest ratio at 25. Also, beyond a genome size of 10$^9$ bp, there appeared to be a plateau in the frequency of TG repeats at $\sim$0.5–2 $\times$ 10$^5$/hap-
Fig. 3.—Dot-blot hybridizations of eukaryotic genomic DNAs. Known amounts of genomic DNA from various eukaryotes, as indicated on the side or top of each column, was blotted onto nitrocellulose and probed for TG elements as in fig. 1 at high stringency. Common names for most of the organisms listed below can be found in table 1. Panel A: 1, HeLa; 2, Chlamydomonas reinhardi; 3, Dictostelium discoideum; 4, Neurospora crassa; 5, Tetrahymena thermophila (total DNA). Panel B: 1, Aurelia victoria; 2, Caenorhabditis elegans; 3, Drosophila melanogaster; 4, Limulus polyphemus; 5, Siren intermedia; 6, Necturus alabamensis; 7, Bufo terrestris; 8, Anguilla rostrata; 9, Bagre marinus; 10, Trachemys scripta; 11, Peromyscus maniculatus; 12, Mus domesticus.

loid genome. Whether the apparent plateau is in fact a true limit to the number of TG elements per genome needs to be tested by assaying organisms with genomes $>10^{10}$ bp.
TG Elements Are Rare in Ciliated Protozoan Genomes

Figure 5 also shows that the two ciliated protozoans, *Oxytricha fallax* and *T. thermophila*, have quite large genomes that contain small numbers of TG elements. Both protozoans are binucleate. The micronucleus functions in sexual reproduction and is transcriptionally inactive. It gives rise to the transcriptionally active macronucleus, which contains the bulk of the DNA of the vegetative cell by a process of DNA amplification and degradation in which the sequence complexity of macronuclear DNA is substantially reduced (Prescott et al. 1973; Yao and Gorovsky 1974). In *Oxytricha* (Prescott et al. 1973), >90% of the micronuclear sequences are lost, leaving DNA in gene-size 2.5-kb fragments, while in *Tetrahymena* (Yao and Gorovsky 1974) the sequence complexity is only reduced by ~15%. The data in table 1 and figure 5 imply that the reduction in sequence complexity in the macronucleus of *Oxytricha* also includes a comparable loss of TG elements. Hence, TG elements are unlikely to play a role in macronuclear sequence selection. Those that are in the macronucleus are probably retained by virtue of occurring in noncoding nucleotide sequences. Although the copy number has not been measured for *Tetrahymena* macro- and micronuclear DNA, a qualitatively similar conclusion can be drawn for *Tetrahymena* from the Southern blot analysis (fig. 2).

The observation that the two protozoans form a separate outlying series with respect to TG elements suggests a distinct line of descent of protozoans within the eukaryotic lineage. It is also noteworthy in this regard that ciliated protozoans also use an atypical genetic code in which the termination codons UAA and UAG specify glutamine (Caron and Meyer 1985; Horowitz and Gorovsky 1985; Preer et al. 1985), a finding that also suggests a very early divergence for this protozoan lineage.

**Myxococcus** and **Halobacter** Lack TG Elements

The bacterium *M. xanthus* and the cellular slime mold *Dictyostelium discoideum* have genome sizes within one order of magnitude of each other (5 × 10^6 and 5.4 × 10^7 bp, respectively) and remarkably similar life cycles (Bonner 1967, p. 43; Kaiser
### Table 1

<table>
<thead>
<tr>
<th>Organism</th>
<th>Genome Size(^a)</th>
<th>Total Copies(^b)</th>
<th>(TG)(_{25})/100 kb(^c)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hela cells (human tumor)</td>
<td>(3 \times 10^9)</td>
<td>48,000</td>
<td>1.6</td>
</tr>
<tr>
<td><em>Mus domesticus</em> (mouse)</td>
<td>(2.7 \times 10^9)</td>
<td>200,000</td>
<td>7.4</td>
</tr>
<tr>
<td><em>Bufo terrestris</em> (toad)</td>
<td>(5 \times 10^9)</td>
<td>18,000</td>
<td>0.36</td>
</tr>
<tr>
<td><em>Anguilla rostrata</em> (eel)</td>
<td>(1.7 \times 10^9)</td>
<td>60,000</td>
<td>3.5</td>
</tr>
<tr>
<td><em>Trachemys scripta</em> (turtle)</td>
<td>(2.3 \times 10^9)</td>
<td>40,000(^d)</td>
<td>1.7</td>
</tr>
<tr>
<td><em>Bagre marinus</em> (catfish)</td>
<td>(2.2 \times 10^9)</td>
<td>6,400</td>
<td>0.29</td>
</tr>
<tr>
<td><em>Limulus polyphemus</em> (horseshoe crab)</td>
<td>(4.6 \times 10^9)</td>
<td>28,000</td>
<td>0.61</td>
</tr>
<tr>
<td><em>Arbacia punctulata</em> (sea urchin)</td>
<td>(7.2 \times 10^8)</td>
<td>95,000</td>
<td>13.2</td>
</tr>
<tr>
<td><em>Aurelia victoria</em> (jellyfish)</td>
<td>(6.9 \times 10^8)</td>
<td>515(^g)</td>
<td>0.08</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em> (fruit fly)</td>
<td>(1.5 \times 10^8)</td>
<td>500</td>
<td>0.33</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em> (nematode)</td>
<td>(8 \times 10^7)</td>
<td>149</td>
<td>0.19</td>
</tr>
<tr>
<td><em>Dictyostelium discoideum</em> (slime mold)</td>
<td>(5.4 \times 10^7)</td>
<td>115</td>
<td>0.21</td>
</tr>
<tr>
<td><em>Neurospora crassa</em> (fungus)</td>
<td>(2.7 \times 10^7)</td>
<td>40</td>
<td>0.15</td>
</tr>
<tr>
<td><em>Oxytricha fallax</em> (protozoan):</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Micronucleus</td>
<td>(5 \times 10^8)</td>
<td>25</td>
<td>0.005</td>
</tr>
<tr>
<td>Macronucleus</td>
<td>(5 \times 10^7)</td>
<td>6</td>
<td>0.012</td>
</tr>
<tr>
<td><em>Tetrahymena thermophila</em> (protozoan)</td>
<td>(2 \times 10^8)</td>
<td>9(^f)</td>
<td>0.005</td>
</tr>
<tr>
<td><em>Chlamydomonas reinhardi</em> (green alga)</td>
<td>(1.2 \times 10^8)</td>
<td>30,000</td>
<td>25</td>
</tr>
<tr>
<td><em>Zea mays</em> (flowering plant)</td>
<td>(5 \times 10^9)</td>
<td>6,500</td>
<td>0.13</td>
</tr>
<tr>
<td>Bacteria (eu- and arch)-</td>
<td>(3.5 \times 10^6)</td>
<td>(0.08)(^g)</td>
<td>...</td>
</tr>
<tr>
<td>Mitochondria (human)</td>
<td>(1.6 \times 10^4)</td>
<td>(0)(^b)</td>
<td>...</td>
</tr>
</tbody>
</table>

\(^a\) Genome size in base pairs came from chemical determinations and/or DNA reassociation kinetics (Fasman 1976, p. 284; Lewin 1980, p. 958).

\(^b\) Reiteration frequency of TG elements per haploid genome estimated from DNA slot blots assuming a length of 50 bp.

\(^c\) Ratio of (TG)\(_{25}\) per 100 kb of haploid genomic DNA.

\(^d\) Taken from chemically determined values for the green, wood, and snapping turtles of 5 pg/diploid nucleus (Fasman 1976, p. 284).

\(^e\) Value taken to be equivalent to that for another jellyfish, *Aurelia aurita* (Lewin 1980, p. 958).

\(^f\) Genome size taken from that reported for *Tetrahymena pyriformis* (Yao et al. 1974).

\(^g\) A low, detectable signal was seen for 1 µg of DNA from *Escherichia coli*, *Myxococcus xanthus*, *Halobacter volcanii*, *Pseudomonas putida*, and *Methanococcus maripaludis* on long exposures, and these (differing by ±10%) have been averaged to estimate copy number for bacteria.

\(^h\) Human mitochondrial DNA was scanned by computer-assisted sequence search for (TG)\(_{n}\), (GT)\(_{n}\), (CA)\(_{n}\), or (AC)\(_{n}\), where \(n = 2-3\).

Under appropriate environmental conditions both aggregate into multicellular fruiting bodies containing 100,000–1,000,000 cells, of which 10% eventually form spores. A difference is that *Dictyostelium* also forms a mobile slug that precedes fruiting-body formation. Despite these similarities, only *Dictyostelium* contains substantial levels of the TG element. The archaeabacterium *Halobacter* also has the eukaryotic-like trait of containing repeated DNA sequences in its genome (Sapienza and Doolittle 1982), yet it does not contain detectable levels of the TG repeat.

**Discussion**

Although for the species analyzed here the copy number of TG repeats generally increased as genome size increased, no simple relationship between genome size and the number of (TG)\(_{25}\)'s was apparent. For example, the ratio of TG repeats/100 kb (excluding bacteria and mitochondria; table 1) varied from 0.005 to 25, or 5,000-fold. With two exceptions (*Chlamydomonas reinhardi* and *Arbacia punctulata*), invertebrates...
had ratios <0.61; also with two exceptions (Bufo terrestris and Bagre marinus), vertebrates had ratios >1.6. The 5,000-fold difference in the ratio occurred between protozoans and Chlamydomonas that are very similar in that they are free-living, flagellated/ciliated, unicellular organisms. It appears, therefore, that the TG element is a dispensable part of the eukaryotic genome.

Nonetheless, the element was found in all eukaryotic genomes assayed. An unsolved problem concerns how the element initially became associated with the eukaryotic genome. One possibility involves the spontaneous nonreparable deamination of 5-methylcytosine to thymine. In higher eukaryotes, the only methylated base is 5-methylcytosine, which occurs most frequently in the dinucleotide CG (reviewed in Razin and Riggs 1980; Doerfler 1983). Several lines of evidence support the hypothesis that deamination of 5-methylcytosine to thymine is responsible for the scarcity of the
CG dinucleotide and for the overabundance of the TG and CA dinucleotides in eu-
karyotic genomes (Bird 1980; McClelland and Ivarie 1982). Hence, a heavily meth-
ylated alternating CG region of DNA might have gradually been converted to TGs in
an ancestral organism. However, in extant organisms, no obvious correlation exists
between an organism's content of 5-methylcytosine and TG elements. *Drosophila* and
*Neurospora*, for instance, contain little if any 5-methylcytosine yet have fairly high
levels of TG repeats in their genomes.

The absence of the repeat from prokaryotic genomes was not unexpected, given
that most prokaryotic genomes lack reiterated sequences similar to those in eu-
karyotes (Sapienza and Doolittle 1982). *Halobacter*, however, is an exception because at least
two species of the bacterium have been shown to contain several families of repeated
sequences that are both clustered and dispersed in the genome (Sapienza and Doolittle
1982). Nonetheless, the simple TG repeat was undetectable in this prokaryote. Why
the short repeat and, perhaps, others of the alternating purine/pyrimidine type are
absent from prokaryotes is unknown. They may not be stable in circular chromosomes
because of recombination- or replication-inducing deletions or deleterious rearrange-
ments. It is known, for example, that some eukaryotic repetitive sequences are not
stable when cloned in *rec* + strains of *Escherichia coli* (Wyman et al. 1985). Furthermore,
the Z DNA–forming sequence poly(dC-dG) is unstable in plasmids in *E. coli* when it
is >30 bp in length (Klysik et al. 1982). Alternating C-G sequences in Z conformation
have also been shown to block elongation of RNA transcripts when cloned downstream
from a promotor. Poly(dT-dG) · poly(dC-dA), however, had no effect on transcription
(Peck and Wang 1985). The foregoing assumes that TG repeats and other simple
sequences arose occasionally during prokaryotic evolution and were subsequently se-
lected against. It will be interesting to determine what happens when such elements
are artificially introduced into mutant strains of *E. coli* defective in various steps of
DNA metabolism.

Finally, the position of TG elements in gene families is not consistently conserved.
Among six members of the human actin-gene family, three TG elements were in
different introns of two genes and absent from the rest (Hamada et al. 1984a), whereas
two nonallelic human gamma-globin genes harbor TG repeats in the same intron
(Slightom et al. 1980). Thus, while TG elements may have arisen by descent in some
gene families, their nonconservation in the actin family implies either that they were
lost during subsequent evolution or that they arose de novo—and may continue to
do so—by some other process(es) (i.e., 5-methylcytosine deamination or mistakes in
DNA replication/repair/recombination).

Acknowledgments

The authors thank the many individuals who generously supplied genomic DNAs
for this analysis and who are listed in the Material and Methods section; Norm Eber-
hardt and John Baxter for the lambda rat prolactin genomic-DNA clone; Marcus
Fechheimer for *Dictyostelium*; and Iain Farrance for comments on the manuscript.
This work was supported by grant CA34066 from the National Cancer Institute
to R.I.

LITERATURE CITED

Bird, A. P. 1980. DNA methylation and the frequency of CpG in animal DNA. Nucleic Acids
Res. 8:1499–1504.


MIESFELD, R., M. KRYSAL, and N. ARNHEIM. 1981. A member of a new repeated sequence family which is conserved throughout eucaryotic evolution is found between human δ and β globin genes. Nucleic Acids Res. 9:5931–5947.


WES M. BROWN, reviewing editor

Received January 16, 1986; revision received March 13, 1986.
Four non-cross-hybridizing, highly repetitive DNA components of the Weddell seal were cloned and used for Southern blot hybridizations in order to clarify pinniped phylogeny. Each of the components was present and possessed the identical fragment length in all pinnipeds, representing true seals, walrus, and sea lions. Three of the four components also hybridized to fragments of the same length in all mustelids except skunk. Limited hybridization also occurred to raccoon DNA. Hybridization to DNA of other terrestrial carnivores (polar bear, dog, cat) was barely recognizable. The fourth component was pinniped specific. The results are compatible with the mustelids and pinnipeds being sister groups. The findings also suggest that the pinnipeds are monophyletic, having separated from mustelid ancestors as one lineage that later differentiated into otariids, odobenids, and phocids. The hybridizations indicated clear differences between the skunk and other mustelids.

Introduction

The application of molecular biological approaches to questions in molecular biology and phylogeny is becoming more widespread. The increasing interest stems from the fact that convergence in the evolution of DNA sequences is unlikely to occur. In population biology, analyses of mitochondrial DNA (mtDNA) are rapidly replacing the use of protein polymorphisms, and mtDNA has been shown to be also suitable for obtaining information on species relationships within a particular genus (Brown 1983; Ferris et al. 1983). Molecular hybridizations based on single-copy DNA have provided valuable information on phyletic relationships at various levels (e.g., O'Brien et al. 1985). A particular advantage of this approach is that long-range relationships can be elucidated. However, as demonstrated by Templeton (1985), who scrutinized hominoid data, considerable care is needed when assessing phyletic affinities in which successive divergence times are small compared to the total elapsed time.

Cladistic characters based on molecular hybridizations that were easily recognized over the range from genus to order would be valuable for studies of evolutionary biology and phylogeny. At least in some instances, highly repetitive DNA appears to fulfill those requirements. So far, however, the use of it has been rather limited, probably because earlier reports proclaimed its species specificity, rather than its usefulness in study of common components in more distantly related taxa.

Studies of DNA in cetaceans have shown that some highly repetitive components can maintain their characteristics through long periods of evolution, yielding unequivocal evidence of phyletic relationships, even at the subordinal level (Árnason 1982b; Árnason et al. 1984; Widegren et al. 1985).

1. Key words: highly repetitive DNA, pinniped evolution.

Address for correspondence and reprints: Dr. Úlfur Árnason, Department of Genetics, University of Lund, Sölvegatan 29, S-223 62 Lund, Sweden.
In the present study we turned our attention to the other large group of marine mammals, namely the pinnipeds, with the following particular aims: (1) to investigate to what extent highly repetitive DNA can be used to elucidate phyletic relationships that have not been conclusively settled by other methods, (2) to answer the question whether the pinnipeds are mono- or diphyletic, and (3) to identify the closest relatives of pinnipeds among terrestrial carnivores. Elucidation of these issues is of particular interest because results based on immunological comparisons suggesting that the pinnipeds are monophyletic (Sarich 1969, 1975) are at variance with most paleontological, anatomical, and zoogeographical studies (McLaren 1960; Mitchell and Tedford 1973; Ray 1976; Tedford 1976; Repenning and Tedford 1977; Janvier 1984). Karyological analyses on pinnipeds have emphasized the prominent karyological uniformity among the pinnipeds (Fay et al. 1967; Ārnason 1974b, 1977, 1982a; Anbinder 1980) but have not answered conclusively the question of their ancestry.

Material and Methods

The pinniped DNAs used in the present study included those of Weddell seal (*Leptonychotes weddelli*), leopard seal (*Hydrurga leptonyx*), crabeater seal (*Lobodon carcinophagus*), bearded seal (*Erignathus barbatus*), hooded seal (*Cystophora cristata*), harbor seal (*Phoca vitulina*), larga seal (*P. largha*), ringed seal (*P. hispida*), harp seal (*P. groenlandica*), walrus (*Odobenus rosomarus*), California sea lion (*Zalophus californianus*), and northern sea lion (*Eumetopias jubatus*). Of the true seals *L. weddelli* H. *leptonyx*, *L. carcinophagus*, *E. barbatus*, and *C. cristata* have 2n = 34 chromosomes; *P. vitulina*, *P. largha*, *P. hispida*, and *P. groenlandica* have 2n = 32. The terrestrial carnivore DNAs included those of the mustelids: otter (*Lutra canadensis*), ferret (*Mustela putorius*), mink (*M. vison*), European badger (*Meles meles*), striped skunk (*Mephitis mephitis*), and skunk (unidentified species). The nonmustelid terrestrial carnivore DNAs were those of raccoon (*Procyon lotor*), polar bear (*Ursus maritimus*), dog (*Canis familiaris*), and cat (*Felis domesticus*).

Preparation and purification of DNA were basically according to the method developed by Marmur (1961). The DNA of the California sea lion was prepared from placenta. This DNA had rather low molecular weight since the placenta was not as fresh as the other tissues used.

Four highly repetitive DNA components were identified in the genome of the Weddell seal after electrophoretic separation of DNA restricted with EcoRI. The lengths of the components were ~2.1 (A), ~1.7 (B), ~0.95 (C), and ~0.7 (D) kbp.

For cloning procedures, DNA was cleaved with EcoRI and separated on 1% preparative agarose gel. After being stained with ethidium bromide, the components were dissected from the gel, electroeluted, ligated into the EcoRI site of the plasmid vector pUC8 (Vieira and Messing 1982), and cloned in *E. coli* K12-JM83. Positive clones were identified by colony hybridization (Grunstein and Wallis 1979).

DNA from the different species included in the comparison was restricted with EcoRI or HaeIII. The digested DNA (1–2 μg) was loaded on 1% agarose gels and separated overnight at low voltage (15–30 V). The gels were stained with ethidium bromide and photographed on a UV transilluminator. The DNA size marker used was pBR322 cleaved with *RsaI*.

Each of the four cloned highly repetitive components of the Weddell seal DNA was labeled with 32P according to the method developed by Rigby et al. (1977) and used for Southern blot hybridizations. The four components used did not cross-hybridize among themselves.
The transfers of DNA to nitrocellulose membranes and the hybridizations were according to Southern (1975). Membranes were prehybridized at 64°C in 3 × SSC (1 × SSC = 0.15 M sodium chloride, 15 mM sodium citrate), 10 × Denhardt’s (1966) solution with 100 μg denatured herring sperm DNA/ml. The same procedures and conditions were used for the hybridizations except that the concentration of Denhardt’s solution was 2×. Membranes were extensively washed in 0.5 × SSC, 0.2% sodium dodecyl sulfate at 64°C.

The hybridized membranes were exposed at −70°C, using Kodak X-Omat RP film and intensifying screens. Time of exposure varied between 4 and 24 h. The autoradiographs were primarily scrutinized with respect to the presence or absence of fragments having the same length as the cloned components.

Results

The results of hybridization using component A are shown in figure 1. Hybridization was primarily registered in an ≈2.1-kbp EcoRI fragment in all pinnipeds and in all mustelids except skunk. In the same materials hybridization also occurred in a 2.7-kbp fragment. In addition, in the pinnipeds hybridization was recorded in larger and smaller fragments, several of which had identical lengths in all species studied. The most conspicuous of these fragments was 3.4 kbp long. Of the phocids included

![Image of autoradiograph](image_url)

**Fig. 1.—Results of EcoRI digestion of DNA of pinnipeds and terrestrial carnivores (above). A, B, C, and D show localization of highly repetitive components of the Weddell seal DNA used for Southern blot hybridizations. Abbreviations in all figs.: L.w., Weddell seal; E.b., bearded seal; C.c., hooded seal; P.v., harbor seal; O.r., walrus; Z.c., California sea lion; E.j. northern sea lion; L.c., otter; M.p., ferret; M.v., mink; M.m., European badger; M.me, striped skunk; P.l., raccoon; U.m., polar bear; C.f., dog; F.c., cat. Marker DNA: pBR322 cleaved with Rsal (size of fragments in kilobases). Hybridization with component A (below) showed that it occurred with identical fragment lengths in all pinnipeds and in all mustelids except the skunk. Low-degree hybridization was registered in the raccoon. The component also hybridized to an ≈2.7-kbp fragment in pinnipeds and mustelids (except skunk). In the pinnipeds hybridization also occurred in various other fragments; the most conspicuous of these was ≈3.4 kbp long. The DNA of Z.c. had lower molecular weight than the DNA of the other species.
in figure 1, the Weddell, bearded, and hooded seals have 2n = 34 chromosomes; the harbor seal has 2n = 32. Hybridization patterns of the Weddell seal and other Antarctic seals (leopard seal, crab eater seal) were indistinguishable, and only the Leptonychotes weddelli results were included. Similarly the harbor seal was used as a representative of the 2n = 32 phocids (ringed seal, large seal, harp seal), which all gave indistinguishable patterns of hybridization. In the 2n = 32 seals and in the hooded seal (2n = 34) distinct hybridization was observed in an ≈1.45-kbp fragment. Hybridization in this area was considerably less in other 2n = 34 species.

Prolonged exposures allowed detection of decreasing degrees of hybridization to an ≈2.1-kbp fragment from the raccoon, striped skunk, and polar bear genomes. No differences were observed between the two skunk materials, and only the results for the striped skunk were included in the figures. Limited hybridization also occurred with cat DNA; in this species, however, the fragment length was different. Hybridization to distinct bands was not registered with dog DNA.

The molecular weight of the California sea lion (Zalophus californianus) DNA was considerably lower than that of other samples used, which presumably accounts (in figs. 1–4) for the hybridization smear in this species and the rather limited representation of larger fragments (in comparison with the included results for other pinniped species).

Component B was well represented in all pinniped genomes, digestion with EcoRI and HaeIII giving probe-hybridizing fragments of identical lengths (fig. 2a and b). In

![Image](image_url)

**Fig. 2.**—Component B was hybridized to DNA cleaved with EcoRI (a) and HaeIII (b), respectively. Digestion with EcoRI gave rise to discrete hybridization fragments in pinnipeds only. In pinnipeds, HaeIII produced the same fragment length as EcoRI. After digestion with HaeIII hybridization to a discrete fragment occurred in both pinnipeds and mustelids (except skunk). Fragment lengths in pinnipeds and mustelids were identical: 1.7 kbp. In the pinnipeds the component also hybridized to some shorter fragments. The ≈0.85-kbp fragment was a characteristic of all Antarctic phocids (Weddell seal, crabeater seal, leopard seal) studied. The pattern of the hooded seal (not shown) was the same as that of other Northern Hemisphere phocids.
Fig. 3.—Results of a hybridization between component C (≈0.95 kbp) and DNA cleaved with EcoRI. The component is pinniped specific, and no hybridization occurred in any DNA of terrestrial carnivores. Fragment lengths were identical in all species.

addition to the distinct hybridization to the EcoRI 1.7-kbp restriction fragment, limited hybridization also occurred in a 2.6-kbp fragment common to all pinnipeds. Outside the pinnipeds, hybridization to discrete fragments was not registered after digestion with EcoRI. Digestion with HaeIII gave identical fragment lengths in pinnipeds and mustelids (except striped skunk), as shown in figure 2b. The degree of hybridization was, however, considerably lower in the mustelids than in the pinnipeds under the more stringent conditions used. This indicates differences in composition in component B between pinniped and mustelid genomes, although the HaeIII fragment length has been preserved. Low degree of hybridization to other terrestrial carnivore DNA was recognized after prolonged exposures. A part of the component was maintained as an ≈1.7-kbp fragment in the raccoon, polar bear, and dog. Restriction with HaeIII revealed the presence of an ≈0.85-kbp fragment in the Weddell seal. This was also a

Fig. 4.—EcoRI cleaved out component D (0.7 kbp) with identical fragment lengths in all pinnipeds and in all mustelids except the skunk. Limited hybridization to a fragment of the same size was recorded in the raccoon. In the pinnipeds, hybridization to fragments of much larger size also occurred. The pattern was almost the same for all pinnipeds.
characteristic of other Antarctic phocids (leopard seal, crabeater seal) studied. The fragment did not occur in Northern Hemisphere phocids.

Component C hybridized with pinniped DNA but not with DNA of terrestrial carnivores (fig. 3). This component is thus a specific pinniped characteristic. Hybridization was primarily registered in an ≈0.95-kbp fragment, but substantial hybridization was also recorded in longer fragments common to all pinnipeds. The lengths of these fragments were 2.2 and 3.9 kbp, respectively. To judge from the degree of hybridization, component C is less frequent in the pinniped genomes than the other components studied.

The results of a hybridization using component D as a labeled probe are shown in figure 4. Hybridization occurred both to the 0.7-kbp fragment and to fragments of much larger size. The component was abundant in all pinnipeds. Outside the pinnipeds it was well represented in the mustelids (except skunk). EcoRI fragment lengths in pinnipeds and mustelids were identical.

The organization of the four components was not studied in any detail beyond the results of the hybridizations shown. It is, however, worth noting that polymeric structures of the components were not observed. Thus the components presumably do not represent the total lengths of interspersed or tandemly organized repeats.

Discussion

As mentioned in the Introduction, analyses of mtDNA and molecular hybridizations based on single-copy DNA have great applicability, although at different levels, for answering phyletic questions. In situations in which the resolution of mtDNA and single-copy DNA analysis is not optimal, highly repetitive DNA can be considered to be a very valuable tool for phyletic comparisons.

The possible functions of highly repetitive DNA are still obscure, but the fact that a particular species may harbor several unrelated highly repetitive components is an important feature that increases the power of resolution in phyletic studies, since some components may be common to the taxa compared whereas others may be unique to a given lineage.

The abundance of highly repetitive DNA makes it relatively easy, using restriction endonucleases, to isolate it from other parts of the genome. This is of particular advantage when the highly repetitive components can be used directly as labeled probes without first cloning them. In hybridizations used for phyletic studies, no significant differences would be expected between cloned and noncloned probes, provided the isolated bands are composed of a single repetitive component. In Southern blot hybridizations, the appearance of discrete autoradiographic bands might facilitate the appreciation by scientists not working in this field of results of molecular hybridizations. Any quantitative interpretation of results from hybridizations using highly repetitive DNA must be made with care, however, since the degree of hybridization will depend not only on the evolutionary distance but also on the copy number of the repeat. In the present study we therefore primarily considered the presence or absence of fragments with specific lengths.

Our results appear to have provided conclusive answers to two questions in pinniped phylogeny: (1) Which are the closest relatives of pinnipeds among terrestrial carnivores? and (2) Have the pinnipeds evolved from their terrestrial ancestors via one or two lineages? These questions have been controversial, for different approaches have given different answers. It was mentioned in the Introduction that most paleontological, anatomical, and zoogeographical information has been interpreted in favor
of a diphyletic origin of pinnipeds. According to that view, phocids share common ancestry with mustelids whereas odobenids and otariids have evolved from ursid ancestors. Immunological studies of albumins and transferrins (Sarich 1969, 1975) have shown a much closer relationship between phocids, odobenids, and otariids than is generally recognized by the advocates of a diphyletic pinniped origin. However, those results have made a limited impression on authors favoring a diphyletic pinniped evolution (e.g., Janvier 1984), probably because the most likely ancestral group among arctoid terrestrial carnivores could not be conclusively identified.

The earliest karyological comparisons of pinnipeds (Fay et al. 1967) could not delineate the terrestrial ancestry of pinnipeds; nor could later studies, based on banded karyotypes, provide a definite answer, although great similarities were found between procyonid and otariid karyotypes (Árnason 1974b, 1977; Wurster-Hill and Gray 1975). The karyological studies clarified the relationships between the $2n = 34$- and $2n = 32$-chromosome phocid karyotypes, however, as well as the relationship between the $2n = 36$- otariid and $2n = 32$-chromosome odobenid karyotypes. Surveying the karyological studies of otariids and phocids, Anbinder (1980) advocated a monophyletic origin of pinnipeds, but he did not link their evolutionary ancestry with that of the mustelids.

In the work reported here, the four unrelated highly repetitive DNA components studied were found in substantial amounts in the phocid, otariid, and odobenid lineages, and their fragment lengths were strikingly well preserved. Two of the components, A and D, were well represented in the mustelid (except skunk) genomes, digestion with EcoRI yielding identical basic fragment lengths in both groups. Component B did not cleave out as a discrete EcoRI fragment in mustelid genomes, but HaeIII gave identical fragment lengths in both pinnipeds and mustelids, although hybridization under the conditions adopted was considerably less in mustelids. Hybridization outside the mustelids was very limited, but the three components occurred with the same fragment length in the raccoon genome. Hybridization to skunk DNA was less than that to raccoon DNA, and hybridization to the other species was still more limited.

Component C had identical fragment length in all pinnipeds. This component was not observed outside the pinnipeds. The specificity of component C may have been acquired in two ways—either by origination of the component in the pinniped lineage or by elimination of the component from the mustelid and other carnivore genomes. In our opinion the first explanation is the more plausible.

Our results indicate that all of the pinnipeds share a common and closer evolutionary ancestry with the mustelids than with any other group included in the comparison. An evolution of phocids from mustelid (lutrine) ancestry has been widely proposed. Evolution of otariids and odobenids from a mustelid rather than ursid-procyonid ancestry also is indicated by the present findings. The Enaliarctinae (Mitchell and Tedford 1973) have been regarded as the link between otarioids and an ursid ancestry. If the Enaliarctinae are the link between otarioids and their terrestrial arctoid ancestors, they must be a link between mustelids and pinnipeds rather than between ursids and pinnipeds. In our opinion the presence of a pinniped-specific highly repetitive DNA component suggests that the pinniped ancestors separated from the mustelids as a common lineage and that the highly repetitive component C proliferated in this lineage prior to the separation of the phocoid and otarioid lineages.

As accounted for by Ray (1976), the evolution of otarioids and phocids is separated with respect to both age and localization of early fossils. It is well substantiated that early otariid evolution took place in areas surrounding the North Pacific Ocean, whereas
Highly Repetitive DNA and Pinniped Phylogeny

phocids evolved in the North Atlantic-Paratethyan region. Our results, which are compatible with a monophyletic origin of pinnipeds, do not necessarily contradict geographically separate origins of otariids and phocids—provided that they arose from a common lineage evolving from the mustelids and that this lineage was represented on both the Pacific and Atlantic sides of present Eurasia (or, alternatively, on both sides of present North America).

It is likely that the evolution of both otariids and phocids was fast in the early stages of marine adaptation; hence evolutionary linkages with mustelids may be difficult to establish by means of paleontological findings. The situation may be similar to that in Cetacea, in which the divergence into odontocetes and mysticetes has been difficult to substantiate with paleontological findings. Monophyletic origin was not established until chromosomal data became available (Arnason 1969, 1974a). Results from later studies on cetacean highly repetitive DNA confirmed the karyological findings (Arnason et al. 1984).

The prominent karyotypic uniformity that characterizes both pinnipeds and cetaceans has been related to limited inbreeding among these mammals in comparison with that in terrestrial mammals (Arnason 1972, 1974a, 1974b, 1982a). The reproductive biology of these animals and their high mobility in an environment without distinct physical barriers are primary factors counteracting the establishment of reproductive isolates in which origination of new karyotypes would be promoted. Wilson et al. (1975) related slow karyotypic evolution to large body size, but this is not a particularly suitable explanation for marine mammals, since cetaceans may differ in size by a factor of $>1,000$ without exhibiting any differences in rates of karyotypic evolution.

Our results show that conservation of the four highly repetitive DNA components in pinnipeds coincides with slow karyotype evolution. However, slow karyotypic evolution is not necessarily linked with slow evolution of highly repetitive DNA. In rorquals, genus *Balaenoptera*, both variable and highly conserved components are present, the variable one having originated the most recently (Arnason and Widegren 1984). In those animals different constraints on the evolution of different highly repetitive components were found, composition and repeat lengths being stringently conserved in some species but not in others. When the phyletic age of the pinnipeds is taken into account, it becomes evident that the characteristics of the four components now studied have been well maintained.

Together with immunological findings (Sarich 1969, 1975), our results indicate a single pinniped evolutionary lineage and that the Pinnipedia should be maintained as a taxonomic unit, rather than placing the Phocidae in the Musteloidea and the Odobenidae and Otariidae in the Otarioidea as proposed by Tedford (1976) and Repenning and Tedford (1977).

It has been suggested (e.g., Ray 1976) that independent invasions from the north gave rise to the different Antarctic phocid species; however, the hybridization of component B to a 0.85-kbp *HaeIII* fragment specific for the different Antarctic phocids suggests that differentiation into species took place in the Antarctic, since different invading populations would be unlikely to carry the same DNA constitution. That argument is not conclusive, however, if the 0.85-kbp fragment is a characteristic of the more ancient *Monachus* as well.

With respect to the three objectives set forth in the Introduction, our results suggest the following conclusions: (1) The earlier findings, based on cetacean comparisons, that highly repetitive DNA can provide answers to phyletic questions, in
some cases even at the subordinal level, are confirmed. (2) The pinniped-specific component suggests that the phocids, otariids, and odobenids evolved as one lineage. (3) Among the terrestrial carnivores the mustelids are the closest relatives to the pinnipeds.

Besides yielding information on the pinnipeds, the molecular hybridizations showed striking differences between the skunk, genus *Mephitis*, and the rest of the mustelid group, composed of the otter, ferret, mink, and European badger. The highly repetitive DNA components common to pinnipeds and the otter, ferret, mink, and European badger were barely recognizable in the skunk. Hybridization was actually greater to the raccoon DNA that it was to the skunk DNA. The results could be interpreted as indicating that the skunk is more distant from the other mustelids, but the definite answer to this question must wait until reciprocal hybridizations, using mustelid probes, have been performed.

Acknowledgments

The authors express their sincere thanks to Drs. K. Benirschke and O. Ryder, San Diego Zoo, San Diego; R. E. Benveniste, NCI, Frederick, Md; W. N. Bonner and R. M. Laws, British Antarctic Survey, Cambridge, England; E. W. Born, Greenland Fishing Research, Copenhagen; F. H. Fay, Institute of Marine Science, Fairbanks, Alaska; and B. O. Röken, Kolmården Zoo, Sweden, for the various samples of tissues and DNA. Prof. F. H. Fay’s and Dr. C. A. Repenning’s valuable comments on the manuscript are also acknowledged. The work was supported by the Swedish Natural Science Research Council (grants B-BU 3655-106 and B-BU 3655-107) and by The Erik Philip-Sörensen Foundation.

LITERATURE CITED


WALTER M. FITCH, reviewing editor

Received January 16, 1986; revision received March 25, 1986.
Differences in Specificity and Catalytic Efficiency between Allozymes of Esterase-4 from *Drosophila mojavensis*

Jan Pen, Albert Schipper, Herman A. H. Rongen, and Jaap J. Beintema

Biochemical Laboratory, Rijksuniversiteit Groningen, The Netherlands

A more than 10-fold difference in the specificity and catalytic efficiency for 1-naphthyl esters was measured between two allozymes of esterase-4 from *Drosophila mojavensis*. This difference is mainly caused by a difference in the affinity for the 1-naphthyl esters. The amino acid compositions of the allozymes are not significantly different, which means that the difference in primary structure is small. Small differences in primary structure generally do not result in such a large increase in catalytic efficiency and such a large shift in substrate specificity as was found in the present study.

Introduction

*Drosophila mojavensis*, which belongs to the repleta group of the genus *Drosophila* (Wasserman 1982), uses as its habitat decaying stems of *Stenocereus gummosus* (agria cactus) in Baja California and of *S. thurberi* (organpipe cactus) on the mainland of northwestern Mexico and the southwestern United States (Heed 1982). Because of their isolated, well-defined habitat, cactophilic *Drosophila* species are, together with the cacti, yeasts, and bacteria on which they live, the subject of several ecological and evolutionary studies (Barker and Starmer 1982).

The esterases of *D. mojavensis* have been studied by Zouros et al. (Zouros 1973; Zouros and Van Delden 1982; Zouros et al. 1982), who demonstrated about six bands with esterase activity on a nondenaturing gel and numbered them according to increasing negative charge. The larva-specific esterase-4 is one of the most polymorphic esterases present in *D. mojavensis*. Esterase-4 is a dimeric enzyme (Zouros and Van Delden 1982) with a subunit molecular weight of 63,000 (Pen et al. 1984, 1986). Esterase-4 and esterase-5 of *D. mojavensis* are the products of a duplicated gene, but the localizations in the body and the ontogenies have become different (Zouros et al. 1982). The common evolutionary origin is clearly demonstrated by their N-terminal sequences, which are 82% identical (Pen et al. 1986).

Normally esterase-4 has a preference for 2- over 1-naphthyl esters, but allozymes have been detected with an altered substrate preference for 1- over 2-naphthyl esters by enzyme-specific histochemical staining on polyacrylamide gels (Zouros and Van Delden 1982). There is no accurate estimate of the relative frequencies of the allozymes of esterase-4 in natural populations. In collected stocks, four alleles coding for enzymes
showing a preference for 1-naphthyl esters on gel had together a frequency of 15%. These four alleles were discovered in stocks that had originated from different localities, and they fall into three distinct electrophoretic classes, which indicates that it is a true polymorphism (Zouros and Van Delden 1982). The four alleles coding for enzymes showing preference for 1-naphthyl esters on gel were found in stocks originating from the peninsula of Baja California and islands in the Gulf of California. They were not found in stocks from the mainland in Mexico or the United States (Zouros and D'Entremont 1980; Zouros and Van Delden 1982). The A426-100α stock, homozygous for an esterase-4 allozyme showing preference for 1-naphthyl esters on gel, originates from Baja California (Zouros and D'Entremont 1980). The BI-86β stock, homozygous for an esterase-4 allozyme showing preference for 2-naphthyl esters on gel, originates from the Sonoran desert (E. Zouros, personal communication).

The allozymes of esterase-4 with a preference for 1-naphthyl esters resemble the bacterial enzymes directed toward novel substrates in vivo found by selection of spontaneous mutants (Hall and Knowles 1976; Hall and Zuzel 1980; Hall 1981; Thurberville and Clarke 1981). Enzyme variation within a species can be used to investigate the importance of natural selection in the initial stages of evolutionary changes of enzymes. Unlike comparisons between species, the effect that alternative biochemical phenotypic characteristics—i.e., catalytic properties—may have on physiological function and fitness can be tested by correlating differences in catalytic efficiency (in $k_{cat}$ and/or $k_{cat}/K_M$; Hall and Koehn 1983; Koehn et al. 1983) or substrate specificity (in $k_{cat}/K_M$; Cornish-Bowden 1979) between allozymes with physiological variation (Koehn et al. 1983). To date, no study fully describes the adaptive molecular mechanism of an enzyme polymorphism (Koehn et al. 1983) and only very few (Place and Powers 1979, 1984; Mane et al. 1983) appreciate the importance of the parameters $k_{cat}$ and $k_{cat}/K_M$ (Koehn et al. 1983). In the present study, the kinetic parameters of two allozymes of esterase-4 were determined, as a first step toward elucidation of the molecular characteristics of this enzyme polymorphism.

**Material and Methods**

**Materials**

Naphthyl esters were obtained from Sigma Chemical Co. (St. Louis). All other reagents were analytical grade products from Merck (Darmstadt, West Germany).

*Drosophila mojavensis* Stocks

The laboratory stocks A426-100α and BI-86β of *D. mojavensis* were used. The stocks have been described by Zouros and Van Delden (1982). Stock A426-100α is homozygous for the 100α allele and stock BI-86β for the 86β allele of esterase-4. The α and β indicate substrate preferences for 1- and 2-naphthyl esters, respectively, as found by enzyme-specific histochemical staining on polyacrylamide gels. The stocks were a gift from E. Zouros (Halifax, Nova Scotia).

Purification of the Allozymes

Larvae, predominantly those in their last instar, were washed with demineralized water and stored at −20 C until used. Both allozymes of esterase-4 were purified from the stocks homozygous for an allozyme by immunoaffinity chromatography followed by anion-exchange high-performance liquid chromatography as described for the 100α allozyme (Pen et al. 1986). Homogeneity in esterase activity was checked using non-
denaturing gel electrophoresis. Following electrophoresis the gels were stained for esterase activity (Pen et al. 1984). Protein homogeneity was checked using sodium dodecyl sulfate–polyacrylamide gel electrophoresis, according to the method of Laemmli (1970), on gels containing 10% polyacrylamide.

**Kinetic Analyses**

Initial rate velocities were determined with a continuous spectrophotometric assay in which the rate of hydrolysis was followed by determining the release of naphthol from the absorbance at 235 nm (Mastropaolo and Yourno 1981). The molar extinction coefficients of 1- and 2-naphthol were determined to be 25,400 and 14,800 M\(^{-1}\) cm\(^{-1}\), respectively. Esterase-4 activity was measured with 4.4–6.3 nM 100α allozyme or 6.8–17.0 nM 86β allozyme at 37 C in 0.04 M sodium phosphate buffer, pH 7.0. The substrate concentrations were evenly spaced from \(\sim 0.2–5 \times K_M\). Enzyme concentrations were determined by amino acid analyses and based on subunit molecular weight. These enzyme concentrations were used to determine \(k_{cat}\) values from the \(V_{max}\) values. Determination of kinetic parameters with the same allozyme from different purifications did give comparable results, indicating that, if any inactive enzyme is present, the amount is similar in different preparations. Reactions were started by adding the enzyme solution into a 1-ml cuvette containing the assay mixture. \(K_M\) and \(V_{max}\) values were determined by fitting the data to a simple Michaelis-Menten model. The data were determined by the nonparametric direct-linear-plot procedure (Eisenthal and Cornish-Bowden 1974). SEs of estimates were calculated with the formulas given by Wilkinson (1961).

**Amino Acid Analyses**

Samples were hydrolyzed in evacuated sealed tubes, using constantly boiling HCl for 24 h at 110 C. Amino acid analyses were performed on a Kontron Liquimatum III analyzer.

**Results**

The purification resulted in allozymes that were homogeneous in terms of esterase activity (not shown) as well as in terms of protein (fig. 1). Additional evidence for homogeneity of esterase-4 as determined by this method has been described before (Pen et al. 1986). Figure 1 also shows that the subunit molecular weights of the two allozymes are identical and between 62,000 and 64,000. The amino acid compositions of the two allozymes show no significant differences (table 2).

The kinetic parameters are given in table 1. The 86β allozyme strongly prefers 2-naphthyl esters over 1-naphthyl esters. It is 10 times more specific for 2- than for 1-naphthyl propionate, and even 37 times more specific for 2- than for 1-naphthyl acetate. These effects are mainly determined by differences in \(K_M\), and only to a small extent by differences in \(k_{cat}\). In contrast with the 86β allozyme, the 100α allozyme does not display as strong a substrate preference: this allozyme is slightly more specific for 1-naphthyl propionate than for 2-naphthyl propionate, because of small differences in \(k_{cat}\). The specificities for 1- and 2-naphthyl acetate are about the same. The main differences between the two allozymes are in the \(K_M\) values for 1-naphthyl esters, whereas the differences in \(k_{cat}\) are small. The \(k_{cat}/K_M\) values of the 100α allozyme for 1-naphthyl acetate and 1-naphthyl propionate are, respectively, 23- and 9-fold higher than those of the 86β allozyme. Compared with the 86β allozyme, the 100α allozyme
Fig. 1.—Sodium dodecyl sulfate-polyacrylamide gel electrophoresis. Electrophoresis was performed on a gel containing 10% polyacrylamide. S = Standard proteins: bovine serum albumin (molecular weight [Mr], 67,000); ovalbumin (Mr 43,000); carbonic anhydrase (Mr 30,000); soybean trypsin inhibitor (Mr 20,100). A = Purified 86β allogymne of esterase-4. B = Purified 100α allogyme of esterase-4.

has acquired the ability to hydrolyze 1-naphthyl esters with the same specificity as it does 2-naphthyl esters.

Discussion

There are several reasons to suppose that specificity for 2-naphthyl esters, like that shown by the 86β allogyme, is the ancestral specificity. First, the corresponding esterase in the sibling species *Drosophila arizonensis* has a specificity for 2-naphthyl esters and no allozymes with an altered specificity were detected (Zouros et al. 1982). Since the two species have the same ancestor (Wasserman 1982), the specificity for 2-naphthyl esters is more likely the ancestral one. Second, no alleles coding for enzymes showing preference for 1-naphthyl esters on gel were detected in stocks originating from the mainland. The absence of this specificity in these stocks would be unlikely if specificity for 1-naphthyl esters were the ancestral condition. Third, allozymes of esterase-5 show a preference for 2-naphthyl esters on gel (Zouros et al. 1982). Since esterase-4 and esterase-5 are the products of a duplicated gene, it is more likely that the specificity for 2-naphthyl esters represents the ancestral situation.

It was found earlier (Pen et al. 1984) that both allozymes have a preference for naphthyl propionate over both naphthyl acetate and naphthyl esters with fatty acids of greater chain length. In this study the preference for naphthyl propionate over naphthyl acetate is confirmed. Moreover, it is shown that the preference is the result
Table 1
Kinetic Parameters of the Esterase-4 Allozymes

<table>
<thead>
<tr>
<th>ALLOZYME AND PARAMETER (units)</th>
<th>1-Naphthyl Acetate</th>
<th>2-Naphthyl Acetate</th>
<th>1-Naphthyl Propionate</th>
<th>2-Naphthyl Propionate</th>
</tr>
</thead>
<tbody>
<tr>
<td>86β:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_M$ (μM)</td>
<td>283.7 (27.1)</td>
<td>14.4 (2.5)</td>
<td>124.0 (7.5)</td>
<td>16.3 (4.5)</td>
</tr>
<tr>
<td>$k_{cat}$ (min$^{-1}$)</td>
<td>367.2 (24)</td>
<td>702.5 (52)</td>
<td>1087 (36)</td>
<td>1382 (136)</td>
</tr>
<tr>
<td>$k_{cat}/K_M$ (μM$^{-1}$ min$^{-1}$)</td>
<td>1.3</td>
<td>48.7</td>
<td>8.8</td>
<td>84.8</td>
</tr>
<tr>
<td>100α:</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$K_M$ (μM)</td>
<td>28.4 (5.1)</td>
<td>58.2 (13.7)</td>
<td>15.6 (2.9)</td>
<td>20.9 (6.3)</td>
</tr>
<tr>
<td>$k_{cat}$ (min$^{-1}$)</td>
<td>832 (75)</td>
<td>925 (105)</td>
<td>1290 (119)</td>
<td>972 (79)</td>
</tr>
<tr>
<td>$k_{cat}/K_M$ (μM$^{-1}$ min$^{-1}$)</td>
<td>29.3</td>
<td>15.8</td>
<td>82.7</td>
<td>46.5</td>
</tr>
</tbody>
</table>

of differences in the $k_{cat}$ as well as in the $K_M$. The differences between the allozymes lie in their specificity for 1-naphthyl esters. Compared to the 86β allozyme, the 100α allozyme has acquired the ability to hydrolyze 1-naphthyl esters because of an increased affinity. In other words, the 100α allozyme has a decreased selectivity for the 2-position on the naphthyl ring. If a highly polymorphic enzyme such as esterase-4 is a "multi-

Table 2
Amino Acid Composition of Esterase-4 Allozymes

<table>
<thead>
<tr>
<th>AMINO ACID</th>
<th>100α$^a$</th>
<th>86β$^b$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asx</td>
<td>12.4 (0.9)</td>
<td>11.4 (1.2)</td>
</tr>
<tr>
<td>Thr</td>
<td>3.7 (0.5)</td>
<td>3.7 (0.9)</td>
</tr>
<tr>
<td>Ser</td>
<td>7.2 (0.8)</td>
<td>6.8 (0.8)</td>
</tr>
<tr>
<td>Glx</td>
<td>12.1 (0.4)</td>
<td>11.1 (1.4)</td>
</tr>
<tr>
<td>Pro</td>
<td>6.6 (0.6)</td>
<td>6.9 (0.9)</td>
</tr>
<tr>
<td>Gly</td>
<td>9.0 (0.4)</td>
<td>8.7 (1.1)</td>
</tr>
<tr>
<td>Ala</td>
<td>8.1 (0.7)</td>
<td>7.9 (0.9)</td>
</tr>
<tr>
<td>Cys</td>
<td>1.3 (0.2)$^c$</td>
<td>1.2 (0.3)$^d$</td>
</tr>
<tr>
<td>Val</td>
<td>4.9 (0.1)</td>
<td>5.3 (0.3)</td>
</tr>
<tr>
<td>Met</td>
<td>1.6 (0.5)</td>
<td>1.5 (0.5)</td>
</tr>
<tr>
<td>Ile</td>
<td>3.4 (0.4)</td>
<td>3.7 (0.6)</td>
</tr>
<tr>
<td>Leu</td>
<td>8.5 (0.2)</td>
<td>8.5 (0.8)</td>
</tr>
<tr>
<td>Tyr</td>
<td>3.5 (0.5)</td>
<td>3.7 (0.4)</td>
</tr>
<tr>
<td>Phe</td>
<td>5.9 (0.4)</td>
<td>5.8 (0.8)</td>
</tr>
<tr>
<td>Lys</td>
<td>5.0 (0.5)</td>
<td>4.5 (0.5)</td>
</tr>
<tr>
<td>His</td>
<td>2.2 (0.1)</td>
<td>2.1 (0.2)</td>
</tr>
<tr>
<td>Arg</td>
<td>5.2 (0.3)</td>
<td>5.3 (0.4)</td>
</tr>
<tr>
<td>Trp</td>
<td>ND</td>
<td>ND</td>
</tr>
</tbody>
</table>

NOTE.—ND = not determined.
$^a$ From four determinations on 0.1–0.7 nmol 100α allozyme.
$^b$ From six determinations on 0.2–0.5 nmol 86β allozyme.
$^c$ From three determinations.
$^d$ From four determinations.
substrate" enzyme, as suggested by Gillespie and Langley (1974), the loss of selectivity
may be an advantage. On the other hand, temporary loss of selectivity may occur
during the transition toward absolute specificity for a new substrate.

The in vivo substrate of esterase-4 in unknown. However, the high specificity
constant of $\sim 10^8 \text{M}^{-1} \text{min}^{-1}$ (table 1) suggests that the in vivo substrate (or substrates!)
must bear some resemblance to the naphthyl esters used in this study. The physiological
significance of the differences in the kinetic properties between the allozymes remains
obscure until the in vivo substrate and function are known. The massive shift in the
specificity suggests a physiological basis and would justify intensive study of the
in vivo substrate and function. A cuticle esterase, such as esterase-4, may hydrolyze
toxic esters produced by the host cactus or by the microorganisms living on these
cacti. Since the in vivo substrate most likely resembles the substrates used in this study
(see above), this may be the clue for the identification of such substances. East (1982)
has found 2-phenylethylacetate in Opuntia rots, the habitat of the closely related D.
buzzatii. The absence of alleles coding for enzymes with a preference for 1-naphthyl
esters in stocks originating from the mainland suggests some correlation with differences
in the environment. The difference in catalytic efficiency and substrate preference
between the esterase-4 allozymes may be correlated with the different cactus habitats
(i.e., cactus chemistry) on mainland and peninsula.

Although the in vivo substrate of esterase-4 and its concentration are unknown,
it is unlikely that the catalytic efficiency is determined by $k_{\text{cat}}$, since saturating substrate
levels are improbable in vivo. For this reason $k_{\text{cat}}/K_M$, which is the rate-determining
parameter at low substrate concentrations, usually is a better indicator of catalytic
efficiency (Hall and Koehn 1983; Hall 1985). This means not only that the substrate
specificity has changed but also that the catalytic efficiency of the 100a allozyme for
1-naphthyl esters has increased compared to that of the 86B allozyme.

The differences in primary structure between allozymes are generally small
(Ramshaw et al. 1979; Retzios and Thatcher 1979). Differences probably arise from
a single or a few point mutations, of which a very limited number result in an amino
acid replacement (Kreitman 1983). The esterase-4 allozymes seem to be no exception.
The subunit molecular weights of the two allozymes are the same (fig. 1), and the
amino acid compositions are not significantly different (table 2). Pilot experiments
with limited proteolysis confirm the resemblance in amino acid sequence (not shown).

Generally, point mutations have effects on the activity of a protein ranging from
slightly negative to fatal—as demonstrated, for instance, by mutant hemoglobins. If
the effects are not negative, then the small differences in primary structure usually
give rise either to no or only small increases in catalytic efficiency or to small changes
in substrate specificity (Place and Powers 1979, 1984; Mane et al. 1983). Differences
in substrate specificity between allozymes are rare, and differences as large as those
found in the present study are exceptional, even with enzymes selected for new sub-
strates (Hall and Knowles 1976; Hall and Zuzel 1980; Hall 1981; Thurberville and

Acknowledgments

We thank Dr. E. Zouros (Halifax, Nova Scotia) for his gift of Drosophila moja-
vensis stocks; the Department of Genetics (Groningen University), and especially Mrs.
M. L. L. Hoeksema, for the excellent maintenance of the stocks; and Dr. R. N. Cam-
pagne for carefully reading the manuscript.
LITERATURE CITED


RICHARD K. KOEHN, reviewing editor

Received November 26, 1985; revision received March 23, 1986.
Information for Contributors

*Molecular Biology and Evolution* is a bimonthly journal devoted to the interdisciplinary science between molecular biology, and evolutionary biology. The journal emphasizes experimental papers, but theoretical papers are also published if they have a solid biological basis. Although this journal is primarily for original papers, review articles and book reviews normally written by solicited authors are also published. Brief discussion and comment on material published in this journal or on issues particularly relevant to readers of this journal will be published as “Letters to the Editor.” Letters that refer to a paper handled by an Associate Editor should be sent to that editor or to the Editor in Chief.

To minimize publication delays, authors should follow the instructions given here and should also provide their telephone numbers.

Submission of Manuscripts

Send manuscripts (one original and two high-quality copies) to the Editor in Chief, Managing Editor, or any Associate Editor (addresses below). Any manuscript or any part of a manuscript which has been published or submitted for publication elsewhere cannot be accepted for publication. Anyone who wishes to write a review article should contact the Managing Editor. Correspondence about book reviews should also be addressed to the Managing Editor. Decision on acceptance of papers will be made as rapidly as possible. Papers that are not suitable to the journal will be returned immediately to authors without detailed review.

After a manuscript is accepted, its author will be requested to sign an agreement transferring copyright to the publisher. No published material may be reproduced or published elsewhere without the written permission of the copyright owner. The journal will not be responsible for the loss of manuscripts at any time.

Publication is taken to imply that the authors are prepared to make available to the public any unpublished sequences on which the paper is based and any clone of cells, DNA, or antibodies used in the experiments reported. This principle also applies to computer programs.

Preparation of Manuscripts

Papers must be written in English and organized in the sequence described below. Each section must be typed double-spaced on heavyweight, nonerasable bond; the page margins should be 1½ inches wide to allow for corrections and manuscript editor’s notes and queries. Special typefaces (e.g., italic or sans serif) should not be used, and right-hand margins should not be justified. Word-processing output on dot matrix printers is acceptable only if it is the quality of the standard typewriter. Handwritten items (e.g., Greek letters) must be identified in the margin. Non-English words must have correct diacritics. Although each major part of the paper (e.g., Literature Cited) must begin on a new page, the pages should be numbered consecutively throughout, beginning with the title page and continuing through the abstract, text, Appendix, Literature Cited, footnotes, tables, and ending with figure legends.

*Title page.*—This page should contain the paper’s title, the names of all authors, the institution(s) at which the research was done, the current affiliations of all authors, the name and address for correspondence, and a footnote on nonstandard abbreviations used, if any (see below). Finally, the title page should also provide a running head (maximum of 50 characters and spaces).

*Abstract.* — The abstract should be a one-manuscript-page factual condensation of the entire paper, including a statement of purpose, a clear description of observations and findings, and a concise presentation of conclusions. It should not assert that the findings are discussed.

*Key words.* — A list of three to six words or phrases should be provided that will accurately index the subject matter of the article.

*Text.* — The text should comprise the following sections: (1) Introduction, (2) Material and Methods, (3) Results, (4) Discussion, and (5) Acknowledgments (if any). Papers should be concise but will not be restricted in length.
All organisms mentioned must be identified by their scientific binomens and underlined. Symbols for genetic loci must also be underlined and should follow the established rules of genetic nomenclature for the various organisms (consult M. Demerec et al., *Genetics* 54 [1966]:61–76). Include the formal IUB name and number of all enzymes mentioned.

Do not use abbreviations for words or phrases used less than five times. Abbreviations used by the *Journal of Biological Chemistry* will be regarded as standard; nonstandard abbreviations should be defined collectively in a footnote.

Mathematical equations must be carefully typewritten: spacing between characters should be correct as typed. It will be assumed that all characters in equations and their counterparts in the text will be set in italics unless the author specifies otherwise the first time a character appears. Equations should be numbered sequentially, in Arabic numerals in parentheses, on the right-hand side of the page.

These and other guidelines can be found in the *Council of Biology Editors Style Manual* (4th ed., 1978). In general, all material should conform to the CBE format. See also recent issues of this journal.

**Terminology.**—A satisfactory interdisciplinary communication requires using words with careful attention to their precise meaning in both disciplines. Authors may use any word they choose provided only that its meaning is clear, consistent, and serves to increase the paper's comprehensibility. The following preferred usages are *not* prescriptive but will be assumed unless authors define them otherwise.

Where the alignments disagree, they are *differences* rather than *changes* since there may have been multiple changes to create a single difference. Differences or changes are *replacements* if the sequences are amino acids, *substitutions* if they are nucleotides. *Mutations* should be restricted to changes before selection has operated. *Homology* must be defined since it has two common meanings: (1) observed similarity and (2) inferred common ancestry. The term *similarity* is preferred for meaning 1 because sequences may have similarity acquired by convergence (analog) rather than retained after divergence (homology). When homology arises via a gene duplication (all or part), it is properly called *paralogy*; when it arises via speciation, it is properly called *orthology*. *Gaps* are introduced into sequences to increase their similarity rather than to optimize similarity (homology), unless an algorithm is employed that guarantees an optimized result according to the way similarity (homology) is defined (e.g., as maximum matches—a third meaning of homology). Similarity should not be asserted to be *significant* unless patently obvious or accompanied by a probability statement and its method of determination (e.g., standard measure, binomial, etc.).

As recommended to the IUB, the preferred single letter code for nucleotide bases including ambiguity is: A = adenine, C = cytosine, G = guanine, T = thymine, U = uracil, R = A/G (purine), Y = C/T (pyrimidine), M = A/C, W = A/T, S = C/G, K = G/T, B = C/G/T (not A), D = A/G/T (not C), H = A/C/T (not G), V = A/C/G (not T), N = X = A/C/G/T (any or unknown). For ambiguous nucleotides, T and U are equivalent.

**Literature Cited.**—Literature in the text should be cited by author and year and, where citation is to a book, the relevant pages thereof. Text citations of two or more works at a time should be given in chronological order. When a paper written by three or more authors is cited, write the name of the first author plus *et al*. The Literature Cited section at the end of the paper should be arranged alphabetically and then chronologically and should contain only works specifically cited in the text. References to papers that have not yet been published will be as for articles (see below), except that "accepted" (along with the journal name) will replace the volume and page numbers. "In press" will not be used. (When such papers include authors of the submitted manuscript, copies of those papers must accompany the submitted manuscript.)

For the style of citations, please note the following examples:

**Journal articles:**

Information for Contributors

Books:


Book chapters:


The abbreviations of periodicals should be those used by the Council of Biology Editors. Periodical titles may also be written out.

Articles should include the name of the reviewing editor (the Editor or Associate Editor with whom author has corresponded) at the end of the Literature Cited section.

Footnotes.—Footnotes should be used sparingly. When necessary, they should be indicated in the text by superscript arabic numerals; the notes themselves should be typed on a page separate from the text. Footnotes to tables are referenced by superscript letters, except for significance levels, which use asterisks; table footnotes should be typed on the same page as the table to which they pertain.

Tables.—Each table must have a brief and self-explanatory title, be numbered with arabic numerals in order of its appearance in the text, and be typed on a separate page. Large, complex tables are discouraged. Guidelines for table format may be found in the CBE Style Manual and the Chicago Manual of Style or may be obtained by writing the editors of this journal.

Legends.—Figure legends should be typed on pages at the end of the manuscript, after tables. Each legend must be descriptive so that the illustration can be understood apart from the text and must define abbreviations used in the illustration.

Illustrations.—Each illustration (figure) should be an original, not a photocopy. Illustrations should be separate and have uniform lettering. They should be numbered consecutively, following the sequence in which they are mentioned in the text. The place where each illustration is to be inserted may be indicated by a circled note in the margin of the typescript. Names of authors, figure number, and an arrow indicating proper orientation should be written lightly in pencil on the back of each figure. Line drawings must be of high quality; typewritten or hand lettering is unacceptable. Photographs should be high-contrast, glossy prints. Magnifications may be indicated by a micron bar or in the legend. Photographs to be reproduced without further reduction must be so marked and may not exceed 4¼ inches wide by 7¾ inches long (122 by 194 mm) in order to fit the journal format. Please keep in mind that these dimensions are maxima. Because of the need for a figure legend, the illustration cannot be the maximum size in both dimensions.

Proofs and Reprints

Offprint order forms will be sent to the author (or in the case of multiple authors, to the senior author) with page proofs. There will be no page charge for publication.
Information for Contributors

Editor in Chief
Walter M. Fitch
Department of Physiological Chemistry
University of Wisconsin—Madison
1300 University Avenue
Madison, Wisconsin 53706

Roy J. Britten
Kerckhoff Marine Laboratory
California Institute of Technology
Corona del Mar, California 92625

Wesley M. Brown
Division of Biological Sciences
University of Michigan
Ann Arbor, Michigan 48109

Richard B. Flavell
Department of Cytogenetics
Plant Breeding Institute
Cambridge CB2 2LQ
England

Managing Editor
Masatoshi Nei
Center for Demographic and Population Genetics
University of Texas Health Science Center at
Houston
P.O. Box 20334
Houston, Texas 77225

Associate Editors
Ken W. Jones
Department of Genetics
University of Edinburgh
West Mains Road
Edinburgh EH9, 3JN Scotland

Richard K. Koehn
Department of Ecology and Evolution
State University of New York at Stony Brook
Stony Brook, New York 11794

Robert K. Selander
Department of Biology
University of Rochester
Rochester, New York 14627
A critical publication for everyone working in molecular biology... Now in its fifth year of publication

DNA

A Journal of Molecular Biology

Editors:
Walter L. Miller, M.D.
and
James Roberts, Ph.D.

FROM LATEST ISSUES

Expression of Rat NADPH—Cytochrome P450 Reductase cDNA in Saccharomyces cerevisiae
By H. Murakami, Y. Yabukasi and H. Ohkawa

A General Method for Retrieving the Components of a Genetically Engineered Fusion Protein
By P.R. Szoka, A.B. Schreiber, and H. Chan

Synthesis of Bovine Prolactin in Escherichia coli
By D.N. Luck, J.K. Ngsee, F.M. Rottman and M. Smith

Isolation and Characterization of the α1-Antitrypsin Gene of Mice.
By K.S. Krauter, B.A. Citron, M.-T. Hsu, D. Powell and J.E. Darnell, J.R.

Construction of a Systematic Set of tRNA Mutants by Ligation of Synthetic Oligonucleotides into Defined Single-Stranded Gaps.
By S.W. Cline, M. Yarus and P. Wier

Binding of the Human Glucocorticoid Receptor to Defined Regions in the Human Growth Hormone and Placental Lactogen Genes.
By P.H. Eliard, M.J. Marchand, G.G. Rousseau, P. Formstecher, M. Mathy-Hartert, A. Belayew, and J.A. Martial

Molecular Cloning of Bovine Viral Diarrhea Viral Sequences.
By A. Renard, C. Guo, D. Schmetz, L. Dagenais, P. Pastoret, D. Dina and J.A. Martial

Hormone-Dependent Covalent Modification and Processing of Human Progestosterone Receptors in the Nucleus.
By K.B. Horowitz, M.D. Francis, and L.L. Wei

A Highly Modular Cloning Vector for the Analysis of Eukaryotic Genes and Gene Regulatory Elements.
By D.S. Pfarr, G. Sathe, and M.E. Reff

Characterization of the Polypeptide Composition of Human Factor VIII C and the Nucleotide Sequence and Expression of the Human Kidney cDNA.
By M.A. Truett et al.

An Ubiquitous Interspersed DNA Sequence Family in An Insect.
By J.Y. Bradfield, J. Locke, and G.R. Wyatt

Cloning and Isolation of Human Cytochrome P450 cDNAs Homologous to Dioxin-Inducible Rabbit mRNAs Encoding P-450 4 and P-450 6.
By L.C. Quattrrochi, S.T. Okino, and U.R. Pendurthi

Structure of the Human Vasoactive Intestinal Polypeptide Gene.

Note to authors: There are no page charges in DNA.
Volume 5, 1986, 6 issues: USA, $135; Overseas Air, $175
Volume 4, 1985, 6 issues: USA, $135; Overseas Air, $175
Volume 3, 1984, 6 issues: USA, $135; Overseas Air, $175
Volume 2, 1983, 4 issues: USA, $135; Overseas Air, $175
Volume 1, 1982, 4 issues: USA, $135; Overseas Air, $175

Send subscription orders to: Mary Ann Liebert, Inc. publishers
157 East 86th Street, New York, N.Y. 10028 • (212) 289-2300
Life Sciences publishes papers concerning the pharmacological aspects of biochemistry, chemistry, endocrinology, immunology, medicinal chemistry, microbiology, molecular biology, pathology, physiology and toxicology. Reviewed on the basis of their scientific merit and the need for rapid publication, these papers cover new information; minireviews on selected facets of a scientific field undergoing rapid change; brief conceptual papers based on original and/or literature data; relevant clinical discoveries; collections of selected papers on current concepts in a particular field; and occasional symposia papers.

Software Survey Section — This new section reports developments in appropriate specialist software.

Subscription Information
Published weekly (Volumes 36 & 37)
Annual subscription (1985)
Two-year rate (1985/86)

US$800.00
US$1520.00

FREE SPECIMEN COPIES AVAILABLE ON REQUEST

Advertising rate card available on request. Back issues and current subscriptions are also available in microform. Prices are subject to change without notice. Journal prices include postage and insurance. Sterling prices are available to UK and Eire customers on request.

Pergamon Press
Headington Hill Hall, Oxford OX3 0BW, UK
Fairview Park, Elmsford, New York 10523, USA

16/6/1984
Evolution as Entropy

Toward a Unified Theory of Biology

Daniel R. Brooks and E. O. Wiley

"By combining recent advances in the physical sciences with some of the novel ideas, techniques, and data of modern biology, this book ambitiously attempts to achieve a new and different kind of evolutionary synthesis. I found it to be challenging, fascinating, infuriating, and provocative, but certainly not dull."

—James H. Brown, Department of Ecology and Evolutionary Biology, University of Arizona

Cloth $25.00 350 pages 82 line drawings
Science and Its Conceptual Foundations series
The research that's led to discovery—chronicled here by the researchers themselves

*Genetic Perspectives in Biology and Medicine*
Edited by Edward D. Garber

This collection of essays, originally published in *Perspectives in Biology and Medicine*—

* covers the dramatic and important developments of the last quarter of this century
* gives first-hand perspectives from the initiators of research and discovery. Among distinguished contributors are Francis Crick, Bruce Wallace, E. L. Tatum, and C. Stern, writing on • the history of genetic research • their fellow researchers • human genetics • the implications and the future of genetic research
* strengthens your foundation in the field, and makes a vital addition to your research library

* PLUS makes an excellent and important addition to textbook lists for courses in—• genetics • history of biology • evolution • biochemistry
  • biomedics • molecular biology

No one in the field should be without this kind of documentation

Order your copies today.

☐ Yes, I want the book that covers the most important developments in biology since Darwin—

Please send me: _______ copies of *Genetic Perspectives in Biology and Medicine* in cloth (ISBN 0-226-28215-5) at $30.00 each.

________ copies of *Genetic Perspectives in Biology and Medicine* in paper (ISBN 0-226-28216-3) at $12.00 each.

Add sales tax: Chicago 8%; elsewhere in Illinois 7%. Outside the Western Hemisphere prices will be slightly higher.

☐ Please send me information on the journal *Perspectives in Biology and Medicine*.

Name ____________________________

Address ____________________________

City ____________________________ State/Country ___________ ZIP _______

Visa or MasterCard accepted. Send charge card information, purchase order, or payment to The University of Chicago Press, Journals Division, Dept. SS7SA, P.O. Box 37005, Chicago, Ill. 60637.
Whether you’re in practice or research, staying ahead can be as easy as 1, 2, 3

1. **The Journal of Infectious Diseases**

- *JID* brings you up-to-date coverage on the wide range of research that influences medical microbiology and your work, covering clinical investigations, experimental medicine, controlled trials, and research from microbiology, immunology, epidemiology, and related disciplines.

- *JID* gives you access to research as soon as it’s available—with lead times of only 6-8 months on most articles, and 2-3 months on communications of unusual interest.

- *JID* provides exposure to the field beyond your usual focus. In addition to original research, *JID* presents concise laboratory and clinical communications, correspondence on items of unusual teaching, clinical, or scientific value, and summaries of workshops.

2. **Reviews of Infectious Diseases**

- *RID* saves you the time of tracking information from several sources. *RID*’s coverage of the most recent literature in medical microbiology includes works on infectious organisms, the immune system and its diseases, and anti-microbial agents; on therapeutic advances; and on results of clinical investigations.

- *RID* makes your exposure to the field complete with review articles, ongoing examinations of work on urgent or timely problems, thematic studies, historical articles, articles on history, and case reports.

3. **Subscribe now.**

*The Journal of Infectious Diseases* is published monthly. Editor: Martha D. Yow

*Reviews of Infectious Diseases* is published bimonthly. Editor: Edward H. Kass

*JID* and *RID* are published by The University of Chicago Press.

<table>
<thead>
<tr>
<th>Individual Rates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Combination JID/RID subscription:</strong> individuals $105.00; institutions $225.00; ASM &amp; IAMS individual members $90.00; students, interns, residents, and fellows (attach copy of ID) $50.00. Outside USA add $22.00.</td>
</tr>
<tr>
<td><strong>Subscription rates for either JID or RID</strong> (be sure to indicate which journal you are ordering): individuals $57.00; institutions $125.00; ASM &amp; IAMS individual members $50.00; students, interns, residents, and fellows (attach copy of ID) $30.00. Outside USA add $12.00 for JID subscription; $10.00 for RID subscription.</td>
</tr>
<tr>
<td>Visa or MasterCard accepted. Send charge card information, purchase order, or payment to The University of Chicago Press, Journals Division, P.O. Box 37005, Dept. SS7SA, Chicago, IL 60637.</td>
</tr>
</tbody>
</table>
Subscribe to BioScience
The biologist's monthly magazine

For professionals and students in every field of biology: Research and policy articles, features, news updates, book reviews, announcements and professional opportunities, letters, and a look at today's computer technologies and new products.

Join AIBS and start your subscription today.

Name ____________________________________________

Address __________________________________________

Charge my □ MC □ VISA Acct. # __________________________

Signature __________________________________________

Rates: □ Individual @ $37.50 per year, □ Family @ $52.00 per year, □ Sustaining @ $55.00 per year (These include $23.00 per year for BioScience), □ Student @ $19.50 per year, □ Emeritus @ $19.50 per year (These include $12.00 per year for BioScience)

Please make checks payable to AIBS. Remit US dollars only. Mail to: AIBS, P.O. Box 27417, Central Station, Washington, DC 20038-7417.

American Institute of Biological Sciences
If the key word is genetics—

behavioral genetics  
clinical genetics and counseling  
immunogenetics  
molecular genetics  
population genetics  
cancer genetics  
developmental genetics  
biological genetics  
cytogenetics  
mutagenesis/somatic  
clonal genetics

—the key research is in

The American Journal of Human Genetics

Established in 1948, the AJHG is one of the leading journals addressing central issues in contemporary genetics. The Journal—which is the official publication of The American Society of Human Genetics—provides a record of research and review relating to heredity in humans and to the application of genetic principles in medicine, psychology, anthropology, and the social sciences as well as in related areas of molecular and cell biology.

Editor: Dr. David E. Comings.

Published bimonthly by The University of Chicago Press. 1-year rates: Individuals $75; Institutions $95. Outside of the U.S.A. postage surcharges vary according to destination. Visa and MasterCard accepted. Mail complete charge card information, payment, or purchase order to The University of Chicago Press, Journals Division, P.O. Box 37005, Chicago, IL 60637. Please note: Subscriptions entered on a calendar-year basis only. Students accepted as members of the society only. For student and all other membership rates, write to the Executive Office, American Society of Human Genetics, P.O. Box 6015, Rockville, MD 20850.

Subscribe to the AJHG.
Do the initials M B E mean something essential to you?

MBE —
Molecular Biology and Evolution —

means first-rate coverage and commitment

— that two areas of science of such increasing importance to each other now have an economical forum for research. In December 1983 MBE made its debut in response to the growing need for more resources specializing in the interactive discoveries and contributions of molecular and evolutionary biology.

— from an editorial staff of leading scientists and researchers in the fields to publish only the most current research — MBE's submission-to-publication time is about 6-7 months.

— to your concerns. MBE is sponsored by the Molecular Biology and Evolution Society and cosponsored by the American Society of Naturalists.

And MBE's specialization means access, in one journal, to work that critically examines the evolutionary significance of macromolecules — papers on such topics as:

- mechanisms of mutational change
- processes of developmental control
- mechanisms of natural selection
- maintenance of genetic polymorphism
- polygenic relationships of organisms
- and theories that integrate these numerous aspects of molecular study

so you save the time you used to spend searching a wide range of disciplines.

MBE means staying apace with the early exponential growth of the field.

SUBSCRIPTION INFORMATION

One-year rates: Individuals $30.00; Institutions $60.00; Students (with copy of ID) $24.00; American Society of Naturalists Individual Members $24.00. Outside USA add $4.00. Visa and MasterCard accepted. To subscribe, send complete charge card information, purchase order, or payment to The University of Chicago Press, Journals Division, Dept. SS6SA, P.O. Box 37005, Chicago, IL 60637.

5/85
Forthcoming

Structure and Regulation of the Anthranilate Synthase Genes in *Pseudomonas aeruginosa*. I. Cloning and Expression in *Escherichia coli*
Irving P. Crawford, Annelie Wilde, Elizabeth M. Yelverton, David Figurski, and Robert W. Hedges

Structure and Regulation of the Anthranilate Synthase Genes in *Pseudomonas aeruginosa*. II. Sequence of trpG Encoding the Glutamine Amidotransferase Subunit
Irving P. Crawford and Lee Eberly

Maintenance of the Cellobiose Utilization Genes of *Escherichia coli* in a Cryptic State
Barry G. Hall, Paul W. Betts, and Maja Kricker

Paenungulata: A Comparison of the Hemoglobin Sequences from Elephant, Hyrax, and Manatee
Traute Kleinschmidt, John Czelusniak, Morris Goodman, and Gerhard Braunitzer

Estimating the Reliability of Evolutionary Trees
David Penny and Michael Hendy

Microcomplement Fixation: A Quantitative Estimator of Protein Evolution
Richard D. Maxson and Linda R. Maxson

Simple Methods for Estimating the Number of Synonymous and Nonsynonymous Nucleotide Substitutions
Masatoshi Nei and Takashi Gojobori

Review of *Population Genetics and Molecular Evolution*, edited by Tomoko Ohta and Kenichi Aoki
Shozo Yokoyama

Richard D. Sage
Concerted Transpositions of Mobile Genetic Elements Coupled with Fitness Changes in *Drosophila melanogaster*  

Sequence of a cDNA for Mouse Thymidylate Synthase Reveals Striking Similarity with the Prokaryotic Enzyme  
Sylvia M. Perryman, Cindy Rossana, Tiliang Deng, Elio F. Vanin, and Lee F. Johnson

Neighboring Base Effects on Substitution Rates in Pseudogenes  
Michael Bulmer

Molecular Evolution of Mammalian Lactate Dehydrogenase-A Genes and Pseudogenes: Association of a Mouse Processed Pseudogene with a B1 Repetitive Sequence  
Kayoko M. Fukasawa, Wen-Hsiung Li, Kiyohito Yagi, Chi-Cheng Luo, and Steven S.-L. Li

The Simple Repeat Poly(dT-dG)·Poly(dC-dA) Common to Eukaryotes Is Absent from Eubacteria and Archaeabacteria and Rare in Protozoans  
Julie Morris, Sidney R. Kushner, and Robert Ivarie

Pinniped Phylogeny Enlightened by Molecular Hybridizations Using Highly Repetitive DNA  
Úlfur Ærnason and Bengt Widegren

Differences in Specificity and Catalytic Efficiency between Allozymes of Esterase-4 from *Drosophila mojavensis*  
Jan Pen, Albert Schipper, Herman A. H. Rongen, and Jaap J. Beintema