Letter to the Editor

Recombination in Animal Mitochondrial DNA

J. Maynard Smith and N. H. Smith

Centre for the Study of Evolution, University of Sussex

Previous attempts to demonstrate, or refute, the occurrence of recombination in mitochondria have used tests designed to measure relatively frequent recombination between rather similar sequences: for example, the homoplasy test (Maynard Smith and Smith 1998) and the regression of linkage disequilibrium between pairs of loci with distance along the chromosome (Awadalla, Eyre-Walker, and Maynard Smith 1999). Recently, Ladoukakis and Zouros (2001; subsequently LZ) have attempted to demonstrate rare recombination events between sequences differing at 10% or more of nucleotides. The aim of this note is, first, to evaluate the statistical support for their conclusion and, second, briefly to discuss its significance.

LZ analyze three published data sets consisting of 1,140-bp sequences of cytochrome *b* from eight individuals of *Rana* and 10 of *Apodemus*, and 366 bp of cytochrome oxidase I from 10 individuals of *Gammarus*: the *Rana* sequences come from different species and their origin is given in table 1. For each set, they first examined a matrix of the variable, informative amino acids (20–23 per set) and identified, by visual inspection, one potential crossover event, involving two parents (*a, c*) and a recombinant (*b*), and two break points identifying a central recombinant piece, in which *b* resembles *c,* and two flanking regions in which *b* resembles *a* (see fig. 1). Thus the proposed event is the transfer of a short central region from a donor, *c*, to a recipient, *a,* yielding sequence *b.*

These putative events are not themselves statistically significant. They are best regarded as hypotheses, whose statistical significance can be tested by examining the full nucleotide sequences. As a measure of recombination, LZ use the ''postrecombination divergence,'' *prd* (see fig. 1): this is a number which would be zero if the two ''parents'' and a recombinant were sequenced before any further variation had arisen and which would then increase by mutation. When all trace of a recombinant has disappeared, the observed value of *prd* will be no greater than the value calculated if the sequence of sites is randomized. LZ performed 1,000 random permutations, retaining the frequencies of each nucleotide at each site but randomizing their sequence. They report that in no case was the value of *prd* for the random sequence as low as that for the actual sequence.

However, it is not clear from their paper just how this calculation was performed. In particular, did their permutations involve all nucleotides, including those coding for amino acid variation and therefore used in formulating the hypothesis they are testing? If these nucleotides were included, then the appropriate significance test would be to randomize a matrix of all the sequences and show that no choice of three sequences and two break points would yield a value of *prd* as statistically significant as that observed; such a test would be difficult to perform. We have therefore repeated their permutations, accepting their identification of the parental and recombinant sequences and proposed break points, but using only sites responsible for synonymous variation in the data set and not used in formulating the hypothesis being tested. The results are shown in table 2. There is convincing evidence of recombination in *Rana* and a strong suggestion of recombination in *Apodemus* but no reason for suspecting recombination in *Gammarus*.

In view of this discrepancy, we decided to use a more direct test of recombination, not dependent on visual inspection of the data. The test used is a modification of the ''maximum chi-square'' method (Maynard Smith 1992). We look for crossovers involving only a single break within the gene (see fig. 2) and use data from variable synonymous sites only, to reduce the likelihood that any observed patterns were caused by selection. For each data set, we compared all pairs of sequences and for each pair all possible crossover points and found the pair of sequences, *a* and *b,* and the break point that maximized the value of chi-square calculated as in figure 2; this value is denoted *maxch.*

To decide whether the event so identified is statistically significant, we generated 1,000 new matrices, maintaining the nucleotide frequency of each site but randomizing their allocation to individuals: thus, we retained the observed polymorphism in the data set but eliminated any linkage disequilibrium. For each matrix we calculated *maxch* as above. The results are given in table 3. Statistical significance depends on whether the observed value of *maxch* is greater than the simulated values.

Table 1
***Rana* Sequences**

Sequence	Accession Number	Species
1 Rb7	AF205088	*R. plancyi*
2 Rn8	AF205087	*R. nigromaculata*
3 Rc6	AF205089	*R. catesbeiana*
4 Ra1	AF205094	*R. amurensis*
5 Rd5	AF205090	*R. dybowskii*
6 Rd4	AF205091	*R. dybowskii*
7 Rr2	AF205093	*R. rugosa*
8 Rr3	AF205092	*R. rugosa*

Key words: recombination, mitochondria, *Rana*, *Apodemus*, *Gammarus.*

Address for correspondence and reprints: J. Maynard Smith, Centre for the Study of Evolution, University of Sussex, BN1 9QL, United Kingdom. E-mail: t.ellis@sussex.ac.uk.

Mol. Biol. Evol. 19(12):2330–2332. 2002

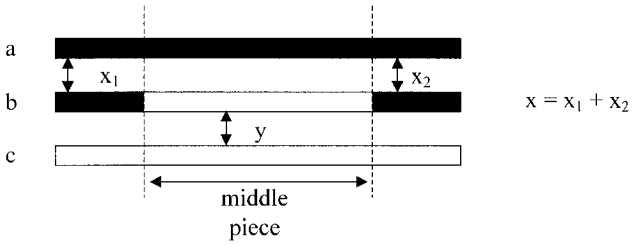

FIG. 1.—Two parental sequences, *a* and *c*, and recombinant, *b*, formed by the insertion of a central region from *c* into *a*. Numbers of nucleotide differences are denoted by *x* and *y*. The ''postrecombinational divergence,'' $prd = x + y$. If the recombinant event is very recent, $x + y = 0$.

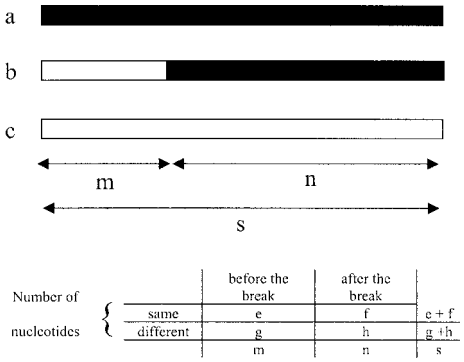

FIG. 2.—A single crossover between sequences *a* and *c*, producing *b*. For *a* versus *b*, or *b* versus *c*, the strength of evidence for a crossover is given by $\chi^2 = (eh - gf)^2 s/mn(e + f)(g + h)$. The expected values of χ^2 are obtained by randomizing the sequences, as explained in the text.

There is no evidence of recombination in either *Gammarus* or *Apodemus*, but there is strong evidence for recombination in *Rana* ($P \ll 0.01$). This supports the conclusion that emerged from our reanalysis of LZ's hypothesis, although the break points and recombinant sequences are not the same as those detected using the method of LZ. The evidence for recombination is illustrated in figure 3. It is not clear which of sequences *a* (Ra1) and *b* (Rc6) is the ''parent'' and which the ''recombinant'': there is no sequence in the data set that is similar to *a* (or to *b*) before the break but different from both *a* and *b* after it, although there are sequences, e.g., sequence *c* (Rr2), that are very different from both *a* and *b* before and after the break.

Both the test suggested by LZ and the maximum chi-square test indicate that, in the *Rana* data, when comparing a pair of sequences, there are regions of similarity and regions of difference along the gene. Such a pattern suggests recombination. As a final confirmation of this conclusion, we applied a modified version of the ''runs test'' suggested by Sawyer (1989) to a matrix of synonymous differences for the three data sets. As before, we found highly significant evidence of ''runs'' in the *Rana* data set but no sign of runs in either *Gammarus* or *Apodemus*.

A pattern of differences of the kind shown in figure 3 strongly suggests a recombination event affecting sequences *a* and *b* (Ra1 and Rc6). Such a pattern could not be generated by clusters of hypervariable sites, or of constrained sites, unless each sequence was subject to a unique set of constraints at synonymous sites. It would require that in sequences *a* and *b*, but not *c*, the sites after the break have a reduced rate of change. This seems implausible, particularly for synonymous sites.

Thus for *Rana*, but not the other genera, there is overwhelming evidence for regions of similarity and difference between sequences for synonymous sites. This is difficult to explain except by recombination. However, there are real difficulties with recombination as an explanation. It is not just that it requires recombination between different ''species'': relatively few recombination events are required to produce the observed patterns. The real difficulty is as follows. The maximum chi-square test reveals seven statistically significant crossovers, each involving two of the sequences numbered 1 to 5 in table 1 and a similar ''break point'' in the range 165 to 181. An examination of the genetic distances between the eight sequences, before and after site 165, reveals that the five sequences are similar to one another after the break point (six of 10 pairwise comparisons differ at less than 20 sites) but very different before the break (nine of 10 pairwise comparisons differ at more than 75 sites). This suggests that, relatively recently, a region of DNA roughly from polymorphic site 165 (site 636) to the end of the available sequence was introduced into each of the five sequences. This seems to require five separate events (or four if one of the sequences was the donor of the DNA). This could perhaps be explained by the spread, by recombination affecting several ''species,'' of a selectively favored region of DNA. The synonymous sites analyzed, although not themselves selected, could have hitch-hiked with the selectively favored amino acid substitutions. It is relevant that there are, in this region, 12 polymorphic amino

Table 2
Test of the Laoukakis and Zouros Hypothesis

	Gammarus	*Apodemus*	*Rana*
Number of sequences	10	10	8
Polymorphic synonymous sites	83	293	260
Ends of 'middle piece'			
All sites	19–288	688–1105	636–863
Polymorphic synonymous sites	4–72	182–289	165–208
Parental sequences[a]	Gf17, Gf1	Ap5, Ag3	Rr2, Rc6
Recombinant sequence[a]	Gf15	As26	Rr3
Trials, out of 1000, with *prd* > observed value	201	11	0

[a] Nomenclature of Laoukakis and Zouros (2001).

Table 3
Maximum Chi-Square Test Using Variable Synonymous Sites Only

	Gammarus	*Apodemus*	*Rana*
Sequences[a]	Gf11, Gf12	As6, Ap5	Ra1, Rc6
Break point[b]	70/288	90/293	167/636
maxch	11.2	12.7	42.1
Trials, out of 100, giving greater *maxch*	33	7	0
Greatest trial *maxch*	15.6	14.1	19.8

[a] Nomenclature of Laoukakis and Zouros (2001).
[b] Position in sequence of variable synonymous sites and of all sites.

acids present in all five sequences but that these are rare in the other three. Members of the genus *Rana* have a number of characteristics that may facilitate the interspecies spread of a selectively favorable region of mitochondrial DNA. These include external fertilization, weak premating isolation and hybrid amphispermy in *R. esculenta* (Graf and Pelaz 1989). However, both in the laboratory and the wild, the progeny of interspecies matings are usually inviable (T. Beebee, personal communication).

To summarize, we can find no evidence for recombination in *Gammarus* and only weak evidence in *Apodemus*, but there is overwhelming evidence for a pattern of similarity and difference at synonymous sites in *Rana*. Although there are difficulties with recombination as an explanation, it is hard to think of any other.

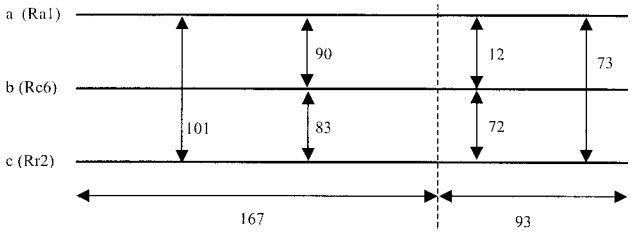

FIG. 3.—A potential recombinant in sequences from *Rana*. Sequences *a* and *b* give a highly significant value of χ^2, for differences before and after polymorphic site 167 (equivalent to nucleotide 636 in the full sequence), as judged by the ''maximum chi-square'' test. However, sequence *c* (chosen because it gives the largest chi-square value compared with either *a* or *b*) is not a plausible second ''parent,'' which should resemble either *a* or *b* before the break.

LITERATURE CITED

AWADALLA, P., A. EYRE-WALKER, and J. MAYNARD SMITH. 1999. Linkage disequilibrium and recombination in hominid mitochondria DNA. Science **286**:2524–2525.

GRAF, J.-D., and M. P. PELAZ. 1989. Evolutionary genetics of the *Rana esculenta* complex. Pp. 289–301 *in*R. M. DAWLEY and J. P. BOGART, eds. Evolution and ecology of unisexual vertebrates. Bulletin 466. New York State Museum, New York.

LADOUKAKIS, E. D., and E. ZOUROS. 2001. Recombination in animal mitochondrial DNA: evidence from published sequences. Mol. Biol. Evol. **18**:2127–2131.

MAYNARD SMITH, J. 1992. Analyzing the mosaic structure of genes. J. Mol. Evol. **35**:126–129.

MAYNARD SMITH, J., and N. H. SMITH. 1998. Detecting recombination from gene trees. Mol. Biol. Evol. **15**:590–599.

SAWYER, S. 1989. Statistical tests for detecting gene conversion. Mol. Biol. Evol. **6**:526–538.

DIETHARD TAUTZ, reviewing editor

Accepted August 6, 2002