Go to CONTENTS

Reprinted from Nature, Vol. 292, No. 5820, pp. 237-239, July 16 1981

Pseudogenes as a paradigm of neutral evolution

Wen-Hsiung Li, Takashi Gojobori & Masatoshi Nei

Center for Demographic and Population Genetics,
The University of Texas at Houston, Houston, Texas 77025, USA

On the neutral mutation hypothesis[1-3], the rate of nucleotide substitution is expected to be higher for functionally less important genes or parts of genes than for functionally more important genes, as the latter would be subject to stronger purifying (negative) selection[2-4]. On the other hand, selectionists believe that most nucleotide substitutions are caused by positive darwinian selection[5,6], in which case the rate of nucleotide substitution in functionally unimportant genes or parts of genes[2,7] is expected to be relatively lower because the mutations in these regions of DNA would not produce any significant selective advantages. Kimura[8] and Jukes[9] have argued that the higher substitution rate observed at the third positions of codons than at the first two positions supports the neutral mutation hypothesis, as most third-position substitutions are synonymous and do not change the amino acids encoded, although others[5,10] have discussed the possibility that third-position substitutions are subject to positive darwinian selection. Recently, Kimura[11] noted that the mouse globin pseudogene, $\psi\alpha 3$, evolved faster than the normal mouse $\alpha 1$ gene, although he did not compute the substitution rate. Here, we present a method of computing the rate of nucleotide substitution for pseudogenes, and report that the three recently discovered pseudogenes show an extremely high rate of nucleotide substitution. As these pseudogenes apparently have no function, this finding strongly supports the neutral mutation hypothesis.

A pseudogene is a DNA segment with high homology with a functional gene but containing nucleotide changes such as frameshift and nonsense mutations that prevent its expression. (Some authors[12] have argued possible functions of pseudogenes, but their arguments are not substantiated.) Pseudogenes seem to have been produced by the nonfunctionalization of duplicate genes. A complete nucleotide sequence is now available for three pseudogenes (mouse $\psi\alpha 3$, human $\psi\alpha 1$ and rabbit $\psi\beta 2$ in the globin gene families)[13-15]. Figure 1 shows a probable evolutionary scheme for mouse pseudogene $\psi\alpha 3$ ($M\psi\alpha 3$), mouse functional gene $\alpha 1$ ($M\alpha 1$)[16] and rabbit functional gene α ($R\alpha$)[17] in which O denotes the point of duplication leading to $M\psi\alpha 3$ and $M\alpha 1$.

Let d_{ABi}, d_{ACi} and d_{BCi} be the numbers of nucleotide substitutions per site at the ith position of codons ($i = 1, 2$ or 3) between

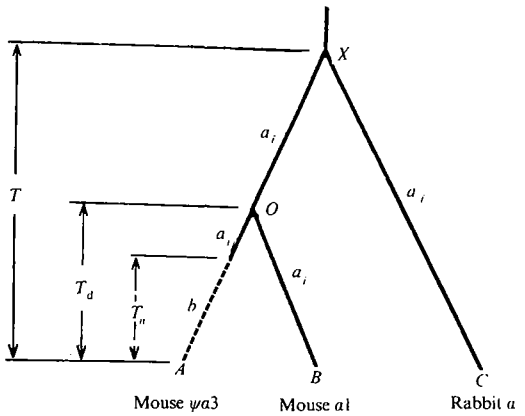

Fig. 1 Plausible phylogenetic tree for mouse $\psi\alpha 3$, mouse $\alpha 1$ and rabbit α. T denotes the divergence time between mouse and rabbit, T_d the time since duplication of mouse $\psi\alpha 3$ and $\alpha 1$, and T_n the time since nonfunctionalization of mouse $\psi\alpha 3$. a_i denotes the rate of nucleotide substitution per site per year at the ith position of codons in the normal globin genes and b the rate of substitution for mouse pseudogene $\psi\alpha 3$. The proportion of nucleotide differences is 51/354 between mouse $\psi\alpha 3$ and $\alpha 1$, 84/423 between mouse $\alpha 1$ and rabbit α, and 91/354 between mouse $\psi\alpha 3$ and rabbit α (see Table 1).

$M\psi\alpha 3$ and $M\alpha 1$, between $M\psi\alpha 3$ and $R\alpha$, and between $M\alpha 1$ and $R\alpha$, respectively, and let l_i, m_i and n_i be the numbers of nucleotide substitutions per site between O and $M\psi\alpha 3$, between O and $M\alpha$ and between O and $R\alpha$, respectively. We then have $d_{ABi} = l_i + m_i$, $d_{ACi} = l_i + n_i$, $d_{BCi} = m_i + n_i$. Therefore, l_i, m_i and n_i can be estimated by

$$l_i = (d_{ABi} + d_{ACi} - d_{Bci})/2 \quad (1a)$$

$$m_i = (d_{ABi} - d_{ACi} + d_{BCi})/2 \quad (1b)$$

$$n_i = (-d_{ABi} + d_{ACi} + d_{BCi})/2 \quad (1c)$$

Pseudogenes often have deletions and insertions. The nucleotides involved in these changes should be eliminated from data analysis, because we are interested only in nucleotide substitution. This can be done by aligning a pseudogene with its homologous functional gene for the coding regions. In our study we followed the alignments given by the original authors[13-15]. In mouse $\psi\alpha 3$, however, we excluded the 30 nucleotides that have been aligned with the nucleotides starting from positions 91 to 120 of the functional mouse $\alpha 1$. In this region there were 21 mismatches, including one sequence of 8 mismatches and another of 7 mismatches. This pattern was very different from that of other parts of the gene and probably occurred through insertion rather than nucleotide substitution. At any rate, after these alignments, we computed the proportion of different nucleotides between homologous genes (Table 1), and from this proportion (p) the total number of nucleotide substitutions per site (d_{ABi}, d_{ACi} or d_{BCi}) was estimated by $d = -(3/4)\ln[1-(4/3)p]$ (ref. 18). For computing d we also used Kimura's formula[11], but the values were not much different.

In Fig. 1 we used rabbit α as a third gene, but human $\alpha 1$ or $\alpha 2$ can also be used for this purpose, as their nucleotide sequences are known[14,19]. We have therefore computed the d values for each of these genes and used the averages of the values to compute l_i, m_i and n_i (Table 2) and the other quantities to be given later. Table 2 also includes the results for human pseudogene $\psi\alpha 1$ and rabbit pseudogene $\psi\beta 2$. It is seen that in the evolution of functional genes (between O and B and between O and C), the third position of codons changes several times faster than the first two positions, which evolve at almost the same rate, but in the line leading to a pseudogene (between O and A) the rates of change for the three positions are about the same and are higher than those between O and B. Clearly, pseudogenes evolve faster than functional genes.

The above computation does not give the substitution rate for pseudogenes. To compute this rate, we must know the time since nonfunctionalization of a pseudogene. Let us again use the three

Table 1 Proportions of nucleotide differences between genes for the three positions of codons

Genes compared	First position	Second position	Third position
$M\psi\alpha 3$ with: $M\alpha 1$	12/118	17/117	22/119
$H\alpha$	19/118	23/117	47/119
$R\alpha$	21/118	22/117	48/119
$H\psi\alpha 1$ with: $H\alpha$	30/133	26/132	42.5/133
$M\alpha 1$	34/133	31/132	49/133
$R\alpha$	39/133	32/132	51/133
$M\alpha 1$ with: $H\alpha$	13/141	14/141	54.5/141
$R\alpha$	16/141	15/141	53/141
$H\alpha$ with: $R\alpha$	18/141	14/141	31/141
$R\psi\beta 2$ with: $R\beta 1$	26/146	19/145	34/145
$H\beta$	31/146	24/145	42/145
$M\beta$	34/146	31/145	43/145
$R\beta 1$ with: $H\beta$	9/146	6.5/146	32/146
$M\beta$	21.5/146	16/146	46/146

$M\psi\alpha 3$ = mouse $\alpha 3$ pseudogene; $M\alpha 1$ = mouse $\alpha 1$; $H\alpha$ = average for human $\alpha 1$ and $\alpha 2$; $R\alpha$ = rabbit α; $H\psi\alpha 1$ = human $\alpha 1$ pseudogene; $R\psi\beta 2$ = rabbit $\beta 2$ pseudogene; $R\beta 1$ = average for rabbit $\beta 1$ allele 1 and allele 2; $H\beta$ = human β, $M\beta$ = mouse β major. The 30 bases near the middle of $M\psi\alpha 3$ are excluded from the comparison (see text).

Table 2 Numbers of nucleotide substitutions per site at the first, second and third positions of codons between *O* and *A*, between *O* and *B* and between *O* and *C*, where *O* is the point of duplication leading to sequences *A* and *B* in Fig. 1

Pseudogene (Sequence *A*)	Sequence *B*	Sequence *C*	Between *O* and *A*			Between *O* and *B*			Between *O* and *C*		
			1	2	3	1	2	3	1	2	3
Mouse $\psi\alpha 3$	Mouse $\alpha 1$	Human α, rabbit α	0.095	0.137	0.125	0.014	0.025	0.088	0.097	0.086	0.446
Human $\psi\alpha 1$	Human α	Mouse $\alpha 1$, rabbit α	0.246	0.205	0.268	0.022	0.024	0.148	0.097	0.083	0.254
Rabbit $\psi\beta 2$	Rabbit $\beta 1$	Human β, mouse β^{maj}	0.184	0.140	0.159	0.020	0.004	0.122	0.086	0.080	0.216

Table 3 Times since gene duplication (T_d), times since nonfunctionalization (T_n) and rates of nucleotide substitution per site per year (b) for pseudogenes mouse $\psi\alpha 3$, human $\psi\alpha 1$ and rabbit $\psi\beta 2$

Pseudogene (Sequence *A*)	Sequence *B*	Sequence *C*	a_1 ($\times 10^{-9}$)	a_2 ($\times 10^{-9}$)	a_3 ($\times 10^{-9}$)	T_d (Myr)	T_n (Myr)	b ($\times 10^{-9}$)
Mouse $\psi\alpha 3$	Mouse $\alpha 1$	Human α, rabbit α	0.69 ± 0.26	0.69 ± 0.26	3.32 ± 0.73	27 ± 6	23 ± 19	5.0 ± 3.2
Human $\psi\alpha 1$	Human α	Mouse $\alpha 1$, rabbit α	0.74 ± 0.27	0.67 ± 0.26	2.51 ± 0.61	49 ± 8	45 ± 37	5.1 ± 3.3
Rabbit $\psi\beta 2$	Rabbit $\beta 1$	Human β, mouse β^{maj}	0.71 ± 0.27	0.51 ± 0.22	2.09 ± 0.51	44 ± 8	44*	3.6*
Average			0.71	0.62	2.64			4.6

a_1, a_2 and a_3 refer to the rates of nucleotide substitutions for the first, second and third positions of codons in the functional genes, respectively.
* We were unable to compute a proper standard error for this estimate, because T_n was assumed to be equal to T_d (see text).

genes in Fig. 1 as an example. In this figure the time (T) since divergence between $M\alpha 1$ and $R\alpha$ is known to be about 80 Myr, but the time (T_d) since duplication of $M\psi\alpha 3$ and $M\alpha 1$ and the time (T_n) since nonfunctionalization of $M\psi\alpha 3$ must be estimated. We estimate these times and the rates of nucleotide substitution for the functional genes and pseudogenes simultaneously. Let a_1, a_2 and a_3 be the rates of nucleotide substitution per site per year at the first, second and third positions of codons in the functional genes, respectively. Once a gene becomes nonfunctional, the rate of nucleotide substitution is expected to be the same for all of the three positions, and we denote the rate by b. From Fig. 1 we obtain

$$d_{ABi} = 2a_iT_d + (b - a_i)T_n \quad (2)$$

$$d_{ACi} = 2a_iT + (b - a_i)T_n \quad (3)$$

$$d_{BCi} = 2a_iT \quad (4)$$

As we know T, a_i can be estimated by $d_{BCi}/(2T)$. We also note

$$y_i \equiv d_{ACi} - d_{BCi} = bT_n - a_iT_n \quad (5)$$

Therefore, from equations (2) and (5), T_d can be estimated by $(\sum d_{ABi} - \sum y_i)/(2\sum a_i)$, where $\sum$ stands for the summation over i.

To estimate T_n and b, we can use equation (5) and apply the standard least-squares method, as there are two unknowns and three equations. However, note that essentially the same results are obtained by the following simple formulae

$$T_n = (y_{12} - y_3)/(a_3 - a_{12}) \quad (6)$$

$$b = (a_3y_{12} - a_{12}y_3)/(y_{12} - y_3) \quad (7)$$

where $y_{12} = (y_1 + y_2)/2$ and $a_{12} = (a_1 + a_2)/2$. Expressing equations (6) and (7) in terms of l_i, m_i and n_i, we have obtained approximate formulae for the standard errors of T_n and b (not shown). In practice, T_n obtained by equation (6) may be larger than T_d. In this case, we set $T_d = T_n$ because by definition $T_n \leq T_d$. When $T_n = T_d$, equation (5) gives $b = (\sum a_iT_d + \sum y_i)/T_d$.

Table 3 shows the results for a_i, T_d, T_n and b. Interestingly, each of a_1, a_2 and a_3 is nearly the same for the three groups of genes studied. However, a_3 is about four times larger than a_1 and a_2. These estimates are similar to those obtained by Kimura[8] and Jukes[9] and support these authors' conclusion that the rate of synonymous substitution is higher than the rate of non-synonymous substitution. The b values in Table 2 indicate that the rate of nucleotide substitution in pseudogenes is even higher than the rate at the third positions of codons in the functional genes. The average rate of 4.6×10^{-9} is one of the highest rates of nucleotide substitutions so far estimated. Only two other estimates are comparable with this value. One is the rate (7.0×10^{-9}) for the synonymous substitution in the C-peptide region of the preproinsulin genes[20], and the other is that (6.2×10^{-9}) estimated from amino acid sequence data for the rapidly evolving residues of fibrinopeptides[21]. These peptides are believed to have no biological function except for holding other polypeptides that will later form a protein. Our finding clearly indicates that functionally less important genes evolve faster than functionally more important genes, and thus supports the neutral mutation hypothesis. Furthermore, comparison of a_3 and b suggests that the third-position substitutions in the globin genes are subject to purifying selection.

Our estimate (T_d) of the time since gene duplication is 27 Myr for mouse $\psi\alpha 3$ and $\alpha 1$, 49 Myr for human $\psi\alpha 1$ and α, and 44 Myr for rabbit $\psi\beta 2$ and $\beta 1$. These estimates are somewhat smaller than those (30, 60, 55 Myr, respectively) obtained by Maniatis and his associates[14,15], mainly because these authors assumed that the substitution rate for the pseudogenes is the same as that for the synonymous changes in functional globin genes. This assumption seems to be incorrect, as the former rate is about twice as great as the latter rate in our study. If our estimates of the times of gene duplication are reliable, the mouse $\psi\alpha 3$ diverged from the $\alpha 1$ gene about 27 Myr ago and became a pseudogene about 4 Myr later. On the other hand, the human $\psi\alpha 1$ was duplicated from the α gene about 49 Myr ago and became nonfunctional about 4 Myr later. The rabbit $\psi\beta 2$ gene became nonfunctional almost immediately after it was duplicated from the $\beta 1$ gene about 44 Myr ago.

We have studied the evolutionary changes of pseudogenes as a paradigm of neutral evolution because in these genes the fate of new mutations in a population is determined almost entirely by genetic drift under the neutral mutation hypothesis. In practice, most genes in the genome would have some biological function and thus be subject to a varying degree of purifying selection. Even in these genes, however, there may be a large number of mutant forms (nucleotide sequences) that are equally functional and fit in adaptation. All these mutant forms will be a source of neutral evolution.

This work was supported by grants from the NIH and NSF.

Note added in proof: After submission of this paper, we learned that Miyata and Yasunaga[22] estimated the rate of nucleotide substitution for mouse $\psi\alpha 3$. Their estimate is higher than ours, because they did not exclude the middle part of the sequence.

Received 17 February; accepted 5 May 1981.

1. Kimura, M. *Nature* **217**, 624–626 (1968).
2. King, J. L. & Jukes, T. H. *Science* **164**, 788–798 (1969).
3. Kimura, M. & Ohta, T. *Proc. natn. Acad. Sci. U.S.A.* **71**, 2848–2852 (1974).
4. Dickerson, R. E. *J. molec. Evol.* **1**, 26–45 (1971).
5. Clarke, B. *Science* **168**, 1009–1011 (1970).
6. Milkman, R. *Trends biochem. Sci.* **1**, N152–N154 (1976).
7. Jukes, T. H. & King, J. L. *Nature* **231**, 114–115 (1971).
8. Kimura, M. *Nature* **267**, 275–276 (1977).
9. Jukes, T. H. *J. molec. Evol.* **11**, 207–209 (1978).

10. Richmond, R. C. *Nature* **225,** 1025–1028 (1970).
11. Kimura, M. *J. molec. Evol.* **16,** 111–120 (1980).
12. Proudfoot, N. J. *Nature* **286,** 840–841 (1980).
13. Nishioka, Y., Leder, A. & Leder, P. *Proc. natn. Acad. Sci. U.S.A.* **77,** 2806–2809 (1980).
14. Proudfoot, N. J. & Maniatis, T. *Cell* **21,** 537–544 (1980).
15. Lacy, E. & Maniatis, T. *Cell* **21,** 545–553 (1980).
16. Nishioka, Y. & Leder, P. *Cell* **18,** 875–882 (1979).
17. Heindell, H. C. *et al. Cell* **15,** 43–54 (1978).
18. Jukes, T. H. & Cantor, C. H. *Mammalian Protein Metabolism* (ed. Munro, H. N.) 21–123 (Academic, New York, 1969).
19. Michelson, A. M. & Orkin, S. H. *Cell* **22,** 371–377 (1980).
20. Perler, F. *et al. Cello* **20,** 555–566 (1980).
21. Kafatos, F. C. *et al. Proc. natn. Acad. Sci. U.S.A.* **74,** 5618–5622 (1977).
22. Miyata, T. & Yasunaga, T. *Proc. natn. Acad. Sci. U.S.A.* **78,** 450–453 (1981).

Printed in Great Britain by Macmillan Production Ltd., Basingstoke, Hampshire