

# Near-neutrality in evolution of genes and gene regulation

Tomoko Ohta\*

National Institute of Genetics, Mishima 411-8540, Japan

This contribution is part of the special series of Inaugural articles by members of the National Academy of Sciences elected on April 30, 2002.

Contributed by Tomoko Ohta, October 16, 2002

**The nearly neutral theory contends that the interaction of drift and selection is important and occurs at various levels, including synonymous and nonsynonymous substitutions in protein coding regions and sequence turnover of regulatory elements. Recent progress of the theory is reviewed, and the interaction between drift and selection is suggested to differ at these different levels. Weak selective force on synonymous changes is stable, whereas its consequence on nonsynonymous changes depends on environmental factors. Selection on differentiation of regulatory elements is even more dependent on environmental factors than on amino acid changes. Of particular significance is the role of drift in the evolution of gene regulation that directly participates in morphological evolution. The range of near neutrality depends on the effective size of the population that is influenced by selected linked loci. In addition to the effective population size, molecular chaperones such as heat shock protein 90 have significant effects on the range of near neutrality.**

Although the neutral and the nearly neutral theories have now been recognized as realistic models to apply to evolutionary changes of genes and proteins, their significance for morphological evolution is still unsettled (1). During the last several years, “slightly advantageous” as well as “slightly deleterious” amino acid substitutions have been noted as important (2–4), and the significance of such weakly selected mutations for morphological evolution needs to be reconsidered. In addition, the molecular basis of gene regulation that is essential for development is being elucidated (5–7). In the evolution of gene regulation, interaction of selection and drift has been suggested to be important (8). The purpose of this article is to review such findings and to expand the concept of near-neutrality.

## Synonymous vs. Nonsynonymous Substitutions

A most efficient way to find how natural selection has worked is to examine the patterns of synonymous and nonsynonymous substitutions. The McDonald and Kreitman test (9) is most popular, and compares the numbers of nonsynonymous and synonymous substitutions within a population and between closely related species. These authors found that the number of fixed amino acid changes at the *Adh* locus between *Drosophila* sibling species is greater than the prediction under the neutral theory, and suggested that the fixed amino acid substitutions were selectively advantageous. Many subsequent studies showed that there were various patterns; i.e., most mitochondria data were characterized by an excess of nonsynonymous within-population (polymorphic) changes, indicating weak negative selection against amino acid changes, whereas nuclear gene data showed more diverse patterns (for a review, see ref. 10).

Recent studies along this line are attempts to examine collections of data of many loci. Smith and Eyre-Walker (11) did a statistical analysis using data of *Drosophila simulans* and *Drosophila yakuba*. Assuming three distinct classes of nonsynonymous changes (deleterious, neutral, and advantageous classes), they estimated that 45% of amino acid substitutions were driven by natural selection. Fay *et al.* (3) compared the ratio of amino

acid to synonymous (A/S) polymorphisms to the ratio A/S for fixed differences between *Drosophila melanogaster* and *D. simulans*, and found that the A/S ratio for fixed difference was twice as large as that for polymorphisms. They attributed the discrepancy between polymorphism and divergence to rapidly evolving proteins, and suggested that it was caused by positive selection.

Another interesting finding is that the A/S pattern differs dramatically between *Drosophila* and *Arabidopsis*. Bustamante *et al.* (4) found evidence for predominantly advantageous amino acid substitutions in *Drosophila* but predominantly deleterious amino acid substitutions in *Arabidopsis*. These authors have suggested that, because of partial selfing of *Arabidopsis*, selection has been inefficient in eliminating slightly deleterious amino acid changing mutations. By assuming that the average value of selection coefficients of mutations at each locus comes from a common normal distribution, they estimated the magnitude of the selection coefficient in terms of the product,  $Ns$ , for polymorphic and fixed amino acid changes, where  $N$  is the effective population size, and  $s$  is the selection coefficient. The average value of  $Ns$  at each locus was found to be mostly negative in *Arabidopsis* and mostly positive in *Drosophila*. The results also show that selection is very weak: the average  $Ns$  falls mostly in the range from  $-2$  to  $+0.5$  for *Arabidopsis* and in the range from  $+0.5$  to  $+3$  for *Drosophila*. In other words, many of these amino acid changes are thought to belong to the nearly neutral class. Remember that the nearly neutral mutations are defined such that their fate in the population depends on both selection and drift (12). Thus, the absolute value of the product,  $Ns$ , should be small, e.g., not larger than 2.

The difference between this nearly neutral model and the traditional selection theory (13) is the magnitude of the selection coefficient. In traditional selection theory, it is assumed that selection is strong enough that the changes in frequency of a mutation depends mainly on selection. Hence the value of  $Ns$  should be at least an order of magnitude larger than the value considered here.

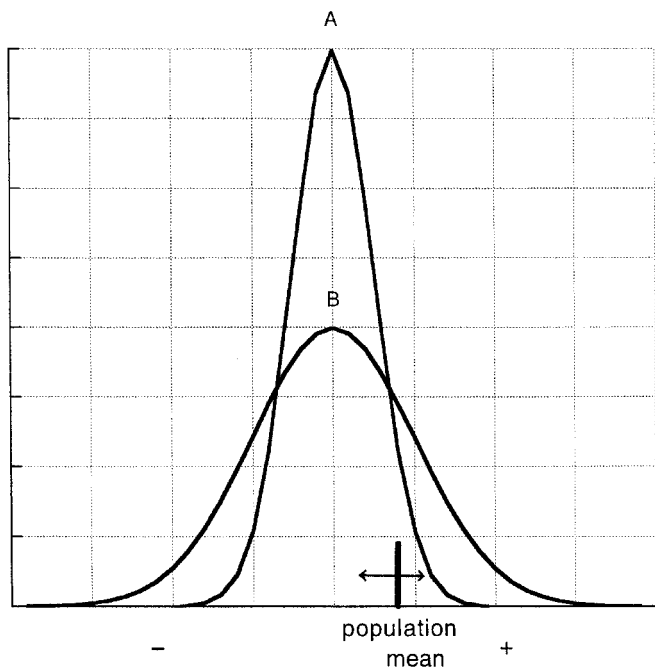
Bustamante *et al.*'s (4) estimates apply to amino acid substitutions that are polymorphic or fixed between the species. Based on these estimates, let us infer the frequency distribution of fitness effects of new mutations that alter an amino acid. Note that the distribution is important for discussions of adaptive vs. nonadaptive evolutionary changes (14).

Let us assume as before that fitness effects are normally distributed. It is natural to suppose that the distribution does not differ much between *Drosophila* and *Arabidopsis*. Bustamante *et al.*'s (4) estimates are thought to be the results of transformation of this distribution via survival probabilities of mutations, and this transformation differs greatly between the two groups of organisms.

A most important factor that needs to be considered is change of population size. Previous analyses are based on equilibrium of mutation-drift-selection, but population size is likely to fluctuate.

Abbreviation: Hsp90, heat shock protein 90.

\*E-mail: tohta@lab.nig.ac.jp.



**Fig. 1.** Schematic diagram showing the frequency distribution of the fitness effects of new mutations. Curve A is for a large population that occupies a heterogeneous environment, and curve B is for a small population that occupies a simple (homogeneous) environment. The population mean moves by drift and selection (modified from ref. 14).

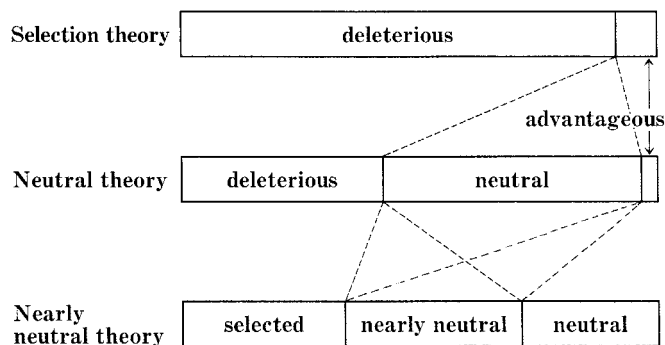
ate, and a nonequilibrium state may be common. In particular, at the time of speciation, population size often goes through a bottleneck (15, 16).

Also the selection coefficients change through space and time (17). Then, at the time of speciation, the magnitudes of drift may get large and the selective force may change. In fact, such changes are related to the shifting-balance theory (18) and to the genetic transience (16).

Let us consider the influence of a bottleneck on the distribution of effects on fitness of mutations. As I have argued (14), when the population size is small, and the population occupies a simple environment, the variance of effects on fitness becomes large as compared with the value at the stable period, even if the mean remains unchanged. This is because the fitness effect is averaged over many environmental conditions if the environment is diverse. The distribution of fitness effects of mutations is shown by the curve B for a small population in Fig. 1, in comparison with that for a stable population (curve A). It is clear that the proportion of nondeleterious mutations, which have some chance of spreading, is much larger for curve B than for curve A. Hence, a large number of nonsynonymous amino acid substitutions between *D. melanogaster* and *D. simulans* (4) may be explained by this nearly neutral model.

It is clear that both drift and selection work simultaneously. The selection intensity may be measured by the standard deviation,  $\sigma$ , of the distribution, whereas the magnitude of drift is measured by the reciprocal of the effective population size,  $1/N$ . So, roughly speaking, if  $\sigma$  is smaller than  $1/N$ , both drift and selection play a role and the population mean moves erratically (19, 20).

On the other hand, the *Arabidopsis* population would be more structured with high inbreeding and limited migration, and therefore the two curves in Fig. 1 do not apply. In fact, the curve at the time of speciation may not differ much from that at the stable period. It should be noted that genetic differentiation among subpopulations is larger in *Arabidopsis* than in *Drosophila* under the present theory.



**Fig. 2.** Schematic diagram showing the proportion of various classes of new mutations, for the selection, the neutral, and the nearly neutral theories (modified from ref. 12).

Another most important factor that influences the range of near neutrality is the activity of molecular chaperones. The best studied case is the heat shock protein 90 (Hsp90) chaperone that facilitates protein folding of many important signal transduction elements. Rutherford and Lindquist (21) reported on drastic phenotypic variation of mutant flies at this locus. Hsp90 is known to help stabilize signal transduction elements acting as a most important chaperone. Proteins with some amino acid changes become unstable unless Hsp90 helps their folding. Mutations at Hsp90 cause various phenotypic variations in *D. melanogaster*. Heterozygous mutant flies at this locus show unusual morphological abnormalities, because Hsp90 activity is not sufficient in heterozygotes. In other words, in normal flies, the effects of various amino acid changes of signal transduction elements are masked by Hsp90, and therefore they become effectively neutral. However in heterozygotes, or at high temperature, some of the amino acid changes reveal themselves as morphological anomalies. Hence, they are conditionally deleterious and belong to the nearly neutral class. Rutherford and Lindquist (21) further argue that such cryptic variations provide mechanisms for large scale morphological evolution. The underlying process is the shifting of adaptive peaks because signal transduction is performed by the interaction of many proteins.

Most mutant combinations are deleterious, but on rare occasions an advantageous or at least nondeleterious mutant combination may appear and spread in the population. Here, the distinction between deleterious and advantageous mutations is not very clear, and the condition for advantageous combination would be met in very limited cases. My previous classification of new mutations is modified so that the nearly neutral class includes intermediates between neutral and advantageous, as well as between neutral and deleterious classes. Fig. 2 presents this concept in a diagram.

Lauter and Doebley (22) have shown that genetic variation for phenotypic traits exists in teosinte even if the variation of these traits is absent under ordinary conditions. The variation was revealed by using a maize-teosinte hybrid made from a cross between an inbred line of maize and a heterozygous teosinte derived from two strains. The traits studied were invariant in the teosinte strains, but showed differences between teosinte and maize. The variation observed in this hybrid is cryptic in teosinte, and these authors argue that such genetic variations are useful for acquiring novel forms in evolution.

Queitsch *et al.* (23) examined Hsp90's effect in *Arabidopsis thaliana* and found that phenotypic plasticity via Hsp90 is even more pronounced in this plant than in flies. So the effect of Hsp90 on phenotype is widespread in animals and plants, and very important for morphological evolution in general.

### Drift vs. Selection in Gene Regulation

It is well known that differential expression of genes is essential for development. In many genes, regulation of their expression is controlled by the interaction between transcription factors and regulatory elements that are located upstream of the coding regions (6). It is thought that gene expression is quantitatively regulated and that the levels of different gene products in a cell are adjusted to achieve a well-balanced state. Any mutations that disturb such a balance are deleterious. The most abundant mutations are nucleotide substitutions in binding sites for transcription factors. Under mutation and selection for maintaining expression patterns, these *cis*-regulatory elements seem to be in constant turnover (6). Binding sites of various transcription factors usually exist in multiples, and some variations are allowed in their sequences. The best studied case is the *stripe2* element of the gene *even-skipped* of *D. melanogaster*. According to Williams *et al.* (24), considerable sequence divergence exists among *Drosophila* species for this element. Ludwig *et al.* (8) constructed a chimera between the *stripe2* element of *D. melanogaster* and that of *Drosophila pseudoobscura*. They compared the expression pattern of a reporter gene downstream from the chimeric element with that downstream from the native element. Whereas the two native elements gave normal expression patterns, the chimeric element gave a different pattern. The authors proposed that stabilizing selection on *eve* expression had led to compensatory turnover of different subregions of the regulatory element. Such changes depend on a series of steps of minor functional modification allowed by drift and selection.

The kinds and quantities of various DNA-binding proteins are regulated and differ in space and time in developing organisms.

Regulatory elements of genes for numerous tissue-specific proteins coevolve with the tissue distribution of DNA-binding proteins. Carroll *et al.* (6) list the mechanisms for evolution of *cis*-regulatory binding sites, i.e., *de novo* formation and evolution from preexisting elements. The latter includes duplication and modification of the elements. All these mechanisms show remarkable flexibility for modifying patterns of tissue-specific gene expression. Because of the short length of DNA binding sites (5–10 base pairs), the chance of *de novo* evolution is not small; however, duplication of preexisting elements may be more frequent (6).

The turnover of regulatory elements may be analogous to nearly neutral amino acid substitutions. Suppose that the distribution of mutant effects on fitness is normally distributed. Note that only the weakly selected class is considered here. More strongly selected mutants exist but are not considered. The two important parameters of the normal distribution are again the mean and the variance of the distribution. For an important protein, i.e., deep in the regulatory network, the mean is negative and the variance is small because few alterations can be tolerated. On the other hand, shortly after the origination of such a protein, the mean would be close to neutrality and the variance would be large. Then the turnover rate is higher for newly recruited loci than for ancient loci.

### Three Levels of Weak Selection

Synonymous substitutions are not completely neutral. For *Drosophila*, Akashi (25) has estimated the magnitude of the selection coefficient is  $Ns \approx 1$ . The selection comes from the efficiency of translation. Because it is ubiquitous, the selection intensity would not depend on environmental factors, and hence differs from the nonsynonymous substitutions. Of course, the selection intensity depends on the kind of proteins, such that

proteins with important function and those produced in large amounts are more selected than others (26). In addition, selection on metabolic efficiency exists on amino acid substitutions (27). This type of selection on nonsynonymous changes as well as that on synonymous substitutions do not depend on environmental factors, and the variance of the distribution of fitness effects would be nil, i.e., selection is global (27). The average selection coefficient differs from locus to locus depending on the requirement of translation or metabolic efficiency.

Contrary to synonymous substitutions, selection on mutations in the regulatory elements of important genes would be strongly influenced by environmental factors, which include both genetic and external factors. As already mentioned, the variance of the distribution of their fitness effects would be larger than the case of nonsynonymous substitutions. Advantageous substitutions in regulatory elements caused by genetic factors are most interesting. They must be responsible for morphological evolution as discussed before. When a new chain of gene expression patterns for transcription factors and signal transduction elements is appearing, many advantageous mutant substitutions are thought to occur simultaneously at the loci participating in the chain. This process is called “recruitment” or “cooption” by developmental biologists (5–7).

How such a chain originates is a very difficult problem, i.e., a module of interacting gene loci would have been constantly tested by natural selection under various genetic and external factors. On very rare occasions, while wandering via mutation and drift under available transcription factors, a module might find its place in a larger gene regulation network. Then positive selection may work on the regulatory elements of the module loci.

In our discussion, the magnitude of drift is very important. Under the common understanding of effective population size, the magnitude of drift may look like rather small and insignificant. However, recent analyses showed that the effect of selected linked loci may greatly reduce the effective population size (28, 29). One should note that even weakly selected loci may have such effects so long as linkage is strong enough. Therefore, the magnitude of drift may not be so small as the apparent population size of a species indicates.

It is important to note the possibility of nearly neutral evolution by gene duplication. It has been repeatedly emphasized that positive selection is needed for the evolution of new functions by gene duplication (30–36). However, it is possible that selection is very weak in this case.

In particular, the evolution of regulatory elements of duplicate genes is likely to be nearly neutral. Lynch and Force (37) have shown that differential expression of duplicate genes, which they call subfunctionalization, is important for preserving duplicate gene copies and for the evolution of new functions.

When a gene under the control of a transcription factor duplicates, its regulatory elements have many ways to differentiate. Because gene duplication may result in imbalance of quantities of gene product, there should be many slightly advantageous and slightly deleterious mutations of the regulatory elements. Nearly neutral nucleotide substitutions and minor duplications and deletions of the regulatory elements of duplicate genes are being tested by natural selection, but in the early stages, their rise and fall in the population are mainly governed by drift (38). Selection pressure seems to be inseparable from the force of drift.

I thank Dr. Dan Hartl for his many valuable suggestions to improve the presentation.

1. Kreitman, M. & Akashi, H. (1995) *Annu. Rev. Ecol. Syst.* **26**, 403–422.

2. Ohta, T. (1997) *J. Mol. Evol.* **44**, S9–S14.

3. Fay, J. C., Wyckoff, G. J. & Wu, C.-I. (2002) *Nature* **415**, 1024–1026.

4. Bustamante, C. D., Nielsen, R., Sawyer, S. A., Olsen, K. M., Purugganan, M. D. & Hartl, D. L. (2002) *Nature* **416**, 531–534.

5. Davidson, E. H. (2001) *Genomic Regulatory Systems* (Academic, San Diego).

6. Carroll, S. B., Grenier, J. K. & Weatherbee, S. D. (2001) *From DNA to Diversity* (Blackwell Scientific, Oxford).

7. Wilkins, A. (2001) *Evolution of Developmental Pathways* (Sinauer, Sunderland, MA).

8. Ludwig, M. Z., Bergman, C., Patel, N. H. & Kreitman, M. (2000) *Nature* **403**, 564–567.
9. McDonald, J. H. & Kreitman, M. (1991) *Nature* **351**, 652–654.
10. Weinreich, D. M. & Rand, D. M. (2000) *Genetics* **156**, 385–399.
11. Smith, N. G. C. & Eyre-Walker, A. (2002) *Nature* **415**, 1022–1024.
12. Ohta, T. (1992) *Annu. Rev. Ecol. Syst.* **23**, 263–286.
13. Gillespie, J. H. (1991) *The Causes of Molecular Evolution* (Oxford Univ. Press, Oxford).
14. Ohta, T. (1972) *J. Mol. Evol.* **1**, 305–314.
15. Mayr, E. (1963) *Animal Species and Evolution* (Belknap, Cambridge, MA).
16. Templeton, A. (1981) *Annu. Rev. Ecol. Syst.* **12**, 23–48.
17. Dykhuizen, D. E. & Hartl, D. L. (1983) *Genetics* **105**, 1–18.
18. Wright, S. (1982) *Annu. Rev. Genet* **16**, 1–19.
19. Ohta, T. & Tachida, H. (1990) *Genetics* **126**, 219–229.
20. Tachida, H. (1991) *Genetics* **128**, 183–192.
21. Rutherford, S. L. & Lindquist, S. (1998) *Nature* **396**, 336–342.
22. Lauter, N. & Doebley, J. (2002) *Genetics* **160**, 333–342.
23. Queitsch, C., Sangster, T. A. & Lindquist, S. (2002) *Nature* **417**, 618–624.
24. Williams, J. A., Paddock, S. W., Vorwerk, K. & Carroll, S. B. (1994) *Nature* **368**, 299–305.
25. Akashi, H. (1995) *Genetics* **139**, 1067–1076.
26. Li, W.-H. (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).
27. Akashi, H. & Gojobori, T. (2002) *Proc. Natl. Acad. Sci. USA* **99**, 3695–3700.
28. Charlesworth, B. & Guttman, D. S. (1996) *J. Genet.* **75**, 49–61.
29. Comeron, J. M. & Kreitman, M. (2002) *Genetics* **161**, 389–410.
30. Ohta, T. (1987) *Genetics* **115**, 207–213.
31. Ohta, T. (1988) *Evolution (Lawrence, Kans.)* **42**, 375–386.
32. Clark, A. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 2950–2954.
33. Walsh, J. B. (1995) *Genetics* **139**, 421–428.
34. Hughes, A. (1999) *Adaptive Evolution of Genes and Genomes* (Oxford Univ. Press, Oxford).
35. Wagner, A. (1999) *J. Evol. Biol.* **12**, 1–16.
36. Long, M., Wang, W. & Zhang, J. (1999) *Gene* **238**, 135–142.
37. Lynch, M. & Force, A. (2000) *Genetics* **154**, 459–473.
38. Ohta, T. (2002) *Genetica*, in press.