Proc. Natl. Acad. Sci. USA
Vol. 90, pp. 7425–7426, August 1993

Commentary

Levels of DNA polymorphism and divergence yield important insights into evolutionary processes

Richard R. Hudson

Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92717

A recent study of Eanes *et al.* (1) exemplifies an approach that promises to go some way toward resolving a long-standing question in population genetics. The question is: To what extent is intra- and interspecific variation in DNA and protein sequences due to molecular noise? The molecular noise hypothesis, more commonly referred to as the neutral theory, maintains that most molecular polymorphism within species and most molecular divergence between species are a consequence of neutral mutations, that is, mutations resulting in functionally equivalent variants, that randomly drift in populations and gradually accumulate over evolutionary time. The alternative to the neutral theory is that most variation within populations is maintained by various forms of balancing selection and/or that the molecular differences between species reflect adaptive changes shaped by natural selection. Ultimately, the issue is a quantitative question; that is, we would like to know how much variation within species and what proportion of changes between species can be attributed to natural selection as opposed to genetic drift of functionally equivalent variants. At present, however, it is not even possible to reject either of the extreme positions (for reviews, see refs. 2 and 3). It should be made clear, however, that both sides in this neutralist/selectionist controversy accept the importance of natural selection as a conservative force, maintaining certain molecular features, by eliminating strongly deleterious mutations quickly after they arise. The conservation of many protein coding sequences and regulatory sequences is strong evidence in support of this aspect of the process.

Over the last 25 years, this controversy has been stewing, with the input of data coming mainly from surveys of enzyme polymorphism and sequencing of proteins. Statistical analysis of enzyme polymorphism data has been useful with respect to this controversy in some cases (for example, see ref. 4) but on the whole has not been very effective in resolving the issue. The power of these analyses is apparently just not up to the task. The clock-like behavior of the amino acid substitution process, as revealed by comparisons of protein sequences from different species, has been used as support for the neutral model. However, careful examination of the data suggests that the substitution process at many loci is much too variable to be compatible with the usual simple versions of the neutral model (3). More complex neutral models have been proposed to explain this property of the substitution process and these models are being debated and contrasted with alternative models that incorporate selection (5–7).

With recent advances in DNA technology, reasonably large surveys of variation at the DNA level are possible. For example, Eanes *et al.* (1) sequenced 32 copies of the *G6pd* gene sampled from natural populations of *Drosophila melanogaster* and 12 copies of the gene from *Drosophila simulans*. Such surveys of DNA variation provide detailed information that can be used in very powerful tests of the neutral model. In particular, the DNA data provide information about the amount of divergence between different copies of the gene sampled from the same species. This divergence within species, or polymorphism, can be directly compared to the level of divergence between copies obtained from different species. In addition, one can compare the levels of divergence within and between species for different kinds of nucleotide sites. For example, for protein coding loci, one can compare the variation at first, second, and third positions of codons. The neutral model makes some very specific predictions about data of this form, and these predictions lead to powerful tests of the model. McDonald and Kreitman (8) have proposed such a test and applied it to data from the *Adh* locus. Eanes *et al.* (1) have now used this test to analyze their sequence data.

The test of McDonald and Kreitman (8) contrasts the level of amino acid divergence and polymorphism with the level of divergence and polymorphism which is silent. (Nucleotide sequence changes that do not alter the protein sequence are referred to as silent changes.) The test is basically a 2×2 test of independence in which the four observations employed are (*i*) the number of "fixed" amino acid replacements between species, (*ii*) the number of "fixed" silent changes between species, (*iii*) the number of amino acid polymorphisms within species, and (*iv*) the number of silent polymorphisms within species. A "fixed" substitution is a mutation that has resulted in all sampled sequences from one species differing from all sampled sequences from another species. (The quotes indicate that "fixed" in this case is only with respect to the particular sample obtained, and another sample might show that what was considered "fixed" in the first sample is a polymorphism in the second sample.) The rows in the 2×2 table would be labeled "fixed" and "polymorphic within species," and the columns would be labeled "amino acid" and "silent." These observed numbers correspond to mutations that have occurred on different parts of the gene genealogy, or gene tree, that represents the history of the sampled genes. To see this, it is useful to consider an example.

Consider a hypothetical data set consisting of the sequences of four copies of a protein coding locus, three copies being obtained by sampling from species A and one copy obtained from species B. [Unlike this example, the data set of Eanes *et al.* (1) consists of multiple copies from each species.] For simplicity, let us assume that no recombination occurs within the locus. In this case, the genealogical history of these sampled sequences can be represented by a simple tree, such as that drawn in Fig. 1. It is important to distinguish between this genealogy, which represents the history of a small sample of copies of the gene, and a tree drawn to represent the phylogenetic history of a set of species. In the gene genealogy, or gene tree, each line represents a single lineage, and any point on a line corresponds to a single copy of the gene that is ancestral to one or more sampled copies. In contrast each line of a phylogenetic tree represents a sequence of many generations of an entire population, and a point on the line corresponds to an entire population.

The variation that is observed in this hypothetical sample is necessarily due to mutations that occurred along the lineages of the gene tree. On the gene tree in

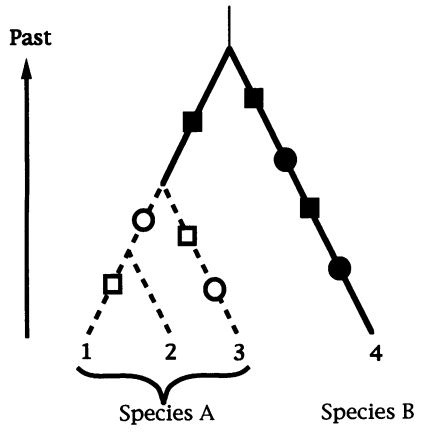

FIG. 1. Example gene tree discussed in the text.

Fig. 1, mutations are indicated by squares and circles. Mutations that occur on the branches drawn with thick solid lines will result in "fixed" differences between sequences. These mutations are indicated by solid squares and circles. Those mutations occurring on dashed branches, indicated by open squares and circles, will result in polymorphism within species A. (We ignore the problems of multiple hits.) The squares, both solid and open, represent amino acid changing mutations; the circles represent silent mutations.

Under the neutral model some statistical properties of these numbers emerge very readily. Let u_a represent the neutral mutation rate per generation for amino acid changes at the locus being considered. That is, each time an offspring is produced, the descendant copy of the gene has probability u_a of differing from its parent by a mutation that changes the amino acid sequence but does not affect the fitness of individuals carrying the mutation. Similarly, let us denote the neutral mutation rate per generation for silent variation by u_s. Under the standard neutral model, these mutation parameters are assumed to be very small and to be constant in time. The occurrence of a mutation in any particular individual is assumed to be independent of mutations that occurred in earlier generations or in other individuals. These assumptions imply that the number of neutral mutations that occur along a lineage, conditional on the length of the lineage, is Poisson distributed with mean equal to the product of the neutral mutation rate and length of the lineage measured in generations. Furthermore, it follows, conditional on the lengths of the branches of the gene tree and the total number of neutral mutations on the gene tree, that the numbers of the four classes of neutral mutations are multinomially distributed. For example, suppose that there are a total of m neutral mutations on the gene tree, that the total length of the thick solid lineages in Fig. 1 is t_b, and that the sum of the lengths of all the branches of the gene tree is t_T, then the expected numbers of "fixed" amino acid differences, amino acid polymorphisms, "fixed" silent differences, and silent polymorphisms are $mp_a p_f$, $mp_a(1 - p_f)$, $m(1 - p_a)p_f$, $m(1 - p_a)(1 - p_f)$, respectively, where $p_a = u_a/(u_a + u_s)$ and $p_f = t_b/t_T$. p_a is the fraction of neutral mutations that are expected to be amino acid changing, and p_f is the fraction of the gene tree that leads to "fixed" differences in the sample. A multinomial distribution with parameters expressible as such products is precisely the standard situation for carrying out a 2×2 test of independence, as suggested by McDonald and Kreitman (8). Eanes *et al.* (1) apply this test to the *G6pd* data and reject neutrality. They suggest that adaptive amino acid substitutions are the cause.

McDonald and Kreitman (8) divide the gene tree into two parts—the between species branches make up one part, and the within species branches make up the other part. There is no reason that one cannot divide up the gene tree into more pieces based on other criteria. In addition, one can clearly define different classes of mutations, other than amino acid changes and silent changes. It is critical, however, for the application of this test, that the different classes of mutation can be assumed to have the same gene tree. Recombination complicates the picture somewhat, because with recombination different parts of the gene can have different gene trees (9). If the different classes of mutations are distributed identically throughout the locus, then recombination does not affect the test, but if the different classes of mutations are spatially segregated in the locus, recombination affects the critical values of the test statistic. Quantitative effects of recombination on this test have not been examined.

The way that McDonald and Kreitman (8) have chosen to divide up the tree into two parts would appear to be particularly appropriate for testing neutral models against many alternative models involving selection. This is because many forms of natural selection will cause departures from neutrality that are in opposite directions for divergence than for polymorphism. Sawyer and Hartl (10) have recently examined the expected patterns of divergence and polymorphism under some relevant models.

One of the strengths of the test of McDonald and Kreitman (8) is that, in contrast to several other tests of neutrality, it does not require that the populations be at statistical equilibrium under the neutral model. The test assumes that the neutral mutation rates have been constant and that there is one gene tree that applies to both classes of variation. No other assumptions are made about the size and shape of the gene tree. This means that things such as population size fluctuations, selection at linked loci, and population subdivision by themselves cannot be used to explain a rejection of the neutral model with this test. On the other hand, we know that the size and shape of gene trees are very sensitive to a variety of population genetic forces, and thus observed gene trees must contain a great deal of information about these forces (9, 11). Some forces will tend to affect all loci in the genome simultaneously, other forces will be tend to be localized in the genome. Comparisons of gene trees from different loci and in closely related species will be useful in discriminating between many hypotheses.

The comparison of levels of divergence and polymorphism for different classes of variation within loci, and for different loci, will provide much insight into population genetic processes. Though such analyses will leave many questions unanswered, they will narrow the range of possibilities considerably, encouraging and complementing other approaches.

1. Eanes, W. F., Kirchner, M. & Yoon, J. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 7475–7479.
2. Kimura, M. (1983) *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, New York).
3. Gillespie, J. H. (1991) *The Causes of Molecular Evolution* (Oxford Univ. Press, New York).
4. Oakeshott, J. B., McKechnie, S. W. & Chambers, G. K. (1984) *Genetics* **63,** 21–29.
5. Gillespie, J. H. (1988) *Genetics* **118,** 385–386.
6. Takahata, N. (1988) *Genetics* **118,** 387–388.
7. Takahata, N. (1991) *Theor. Popul. Biol.* **39,** 329–344.
8. McDonald, J. H. & Kreitman, M. (1991) *Nature (London)* **351,** 652–654.
9. Hudson, R. R. (1990) in *Oxford Surveys in Evolutionary Biology*, eds. Futuyma, D. & Antonovics, J. (Oxford Univ. Press, Oxford), pp. 1–44.
10. Sawyer, S. A. & Hartl, D. L. (1992) *Genetics* **132,** 1161–1176.
11. Takahata, N. & Nei, M. (1990) *Genetics* **124,** 967–978.