

THE NEUTRAL THEORY OF MOLECULAR EVOLUTION

Introduction

I didn't make a big deal of it in what we just went over, but in deriving the Jukes-Cantor equation I used the phrase "substitution rate" instead of the phrase "mutation rate." As a preface to what is about to follow, let me explain the difference.

- *Mutation rate* refers to the rate at which changes are incorporated into a nucleotide sequence during the process of replication, i.e., the probability that an allele differs from the copy of that in its parent from which it was derived. *Mutation rate* refers to the rate at which mutations arise.
- An allele substitution occurs when a newly arisen allele is incorporated into a population, e.g., when a newly arisen allele becomes fixed in a population. *Substitution rate* refers to the rate at which allele substitutions occur.

Mutation rates and substitution rates are obviously related related—substitutions can't happen unless mutations occur, after all—, but it's important to remember that they refer to different processes.

Early empirical observations

By the early 1960s amino acid sequences of hemoglobins and cytochrome *c* for many mammals had been determined. When the sequences were compared, investigators began to notice that the number of amino acid differences between different pairs of mammals seemed to be roughly proportional to the time since they had diverged from one another, as inferred from the fossil record. Zuckerkandl and Pauling [8] proposed the *molecular clock hypothesis* to explain these results. Specifically, they proposed that there was a constant rate of amino acid substitution over time. Sarich and Wilson [6, 7] used the molecular clock hypothesis to propose that humans and apes diverged approximately 5 million years ago. While that

proposal may not seem particularly controversial now, it generated enormous controversy at the time, because at the time many paleoanthropologists interpreted the evidence to indicate humans diverged from apes as much as 30 million years ago.

One year after Zuckerkandl and Pauling's paper, Harris [1] and Hubby and Lewontin [2, 5] showed that protein electrophoresis could be used to reveal surprising amounts of genetic variability within populations. Harris studied 10 loci in human populations, found three of them to be polymorphic, and identified one locus with three alleles. Hubby and Lewontin studied 18 loci in *Drosophila pseudoobscura*, found seven to be polymorphic, and five that had three or more alleles.

Both sets of observations posed real challenges for evolutionary geneticists. It was difficult to imagine an evolutionary mechanism that could produce a constant rate of substitution. It was similarly difficult to imagine that natural selection could maintain so much polymorphism within populations. The "cost of selection," as Haldane called it would simply be too high.

Neutral mutations

Kimura [3] and King and Jukes [4] proposed a way to solve both empirical problems. If the vast majority of amino acid substitutions are selectively neutral, then substitutions will occur at approximately a constant rate (assuming that mutation rates don't vary over time) and it will be easy to maintain lots of polymorphism within populations because there will be no cost of selection. I'll develop both of those points in a bit more detail in just a moment, but let me first be precise about what the neutral theory of molecular evolution actually proposes. More specifically, let me first be precise about what it does *not* propose. I'll do so specifically in the context of protein evolution for now, although we'll broaden the scope later.

- *The neutral theory asserts that alternative alleles at variable protein loci are selectively neutral.* This does *not* mean that the locus is unimportant, only that the alternative alleles found at this locus are selectively neutral.
 - Glucose-phosphate isomerase is an essential enzyme. It catalyzes the first step of glycolysis, the conversion of glucose-6-phosphate into fructose-6-phosphate.
 - Natural populations of many, perhaps most, populations of plants and animals are polymorphic at this locus, i.e., they have two or more alleles with different amino acid sequences.
 - The neutral theory asserts that the alternative alleles are selectively neutral.

- By *selectively neutral* we do *not* mean that the alternative alleles have no effect on physiology or fitness. We mean that the selection among different genotypes at this locus is sufficiently weak that the pattern of variation is determined by the interaction of mutation, drift, mating system, and migration. This is roughly equivalent to saying that $N_e s < 1$, where N_e is the effective population size and s is the selection coefficient on alleles at this locus.
 - Experiments in *Colias* butterflies, and other organisms have shown that different electrophoretic variants of GPI have different enzymatic capabilities and different thermal stabilities. In some cases, these differences have been related to differences in individual performance.
 - If populations of *Colias* are large and the differences in fitness associated with differences in genotype are large, i.e., if $N_e s > 1$, then selection plays a predominant role in determining patterns of diversity at this locus, i.e., the neutral theory of molecular evolution would not apply.
 - If populations of *Colias* are small or the differences in fitness associated with differences in genotype are small, or both, then drift plays a predominant role in determining patterns of diversity at this locus, i.e., the neutral theory of molecular evolution applies.

In short, the neutral theory of molecular really asserts only that observed amino acid substitutions and polymorphisms are *effectively* neutral, not that the loci involved are unimportant or that allelic differences at those loci have no effect on fitness.

The rate of molecular evolution

We're now going to calculate the rate of molecular evolution, i.e., the rate of allelic substitution, under the hypothesis that mutations are selectively neutral. To get that rate we need two things: the rate at which new mutations occur and the probability with which new mutations are fixed. In a word equation

$$\begin{aligned} \# \text{ of substitutions/generation} &= (\# \text{ of mutations/generation}) \times (\text{probability of fixation}) \\ \lambda &= \mu_0 p_0 \quad . \end{aligned}$$

Surprisingly,¹ it's pretty easy to calculate both μ_0 and p_0 from first principles.

¹Or perhaps not.

In a diploid population of size N , there are $2N$ gametes. The probability that any one of them mutates is just the mutation rate, μ , so

$$\mu_0 = 2N\mu \quad . \quad (1)$$

To calculate the probability of fixation, we have to say something about the dynamics of alleles in populations. Let's suppose that we're dealing with a single population, to keep things simple. Now, you have to remember a little of what you learned about the properties of genetic drift. If the current frequency of an allele is p_0 , what's the probability that it is eventually fixed? p_0 . When a new mutation occurs there's only one copy of it,² so the frequency of a newly arisen mutation is $1/2N$ and

$$p_0 = \frac{1}{2N} \quad . \quad (2)$$

Putting (1) and (2) together we find

$$\begin{aligned} \lambda &= \mu_0 p_0 \\ &= (2N\mu) \left(\frac{1}{2N} \right) \\ &= \mu \quad . \end{aligned}$$

In other words, if mutations are selectively neutral, the substitution rate is equal to the mutation rate. Since mutation rates are (mostly) governed by physical factors that remain relatively constant, mutation rates should remain constant, implying that substitution rates should remain constant if substitutions are selectively neutral. In short, if mutations are selectively neutral, we expect a molecular clock.

Diversity in populations

Protein-coding genes consist of hundreds or thousands of nucleotides, each of which could mutate to one of three other nucleotides.³ That's not an infinite number of possibilities, but it's pretty large.⁴ It suggests that we could treat every mutation that occurs as if it were completely new, a mutation that has never been seen before and will never be seen again. Does that description ring any bells? Does the infinite alleles model sound familiar? It should, because it exactly fits the situation I've just described.

²By definition. It's new.

³Why three when there are four nucleotides? Because if the nucleotide at a certain position is an A, for example, it can only *change* to a C, G, or T.

⁴If a protein consists of 400 amino acids, that's 1200 nucleotides. There are $4^{1200} \approx 10^{720}$ different sequences that are 1200 nucleotides long.

Having remembered that this situation is well described by the infinite alleles model, I'm sure you'll also remember that we can calculate the equilibrium inbreeding coefficient for the infinite alleles model, i.e.,

$$f = \frac{1}{4N_e\mu + 1} \quad .$$

What's important about this for our purposes, is that to the extent that the infinite alleles model is appropriate for molecular data, then f is the frequency of homozygotes we should see in populations and $1 - f$ is the frequency of heterozygotes. So in large populations we should find more diversity than in small ones, which is roughly what we do find. Notice, however, that here we're talking about heterozygosity at individual nucleotide positions,⁵ not heterozygosity of haplotypes.

Conclusions

In broad outline then, the neutral theory does a pretty good job of dealing with at least some types of molecular data. I'm sure that some of you are already thinking, "But what about third codon positions *versus* first and second?" or "What about the observation that histone loci evolve much more slowly than interferons or MHC loci?" Those are good questions, and those are where we're going next. As we'll see, molecular evolutionists have elaborated the framework extensively⁶ in the last thirty years, but these basic principles underlie every investigation that's conducted. That's why I wanted to spend a fair amount of time going over the logic and consequences. Besides, it's a rare case in population genetics where the fundamental mathematics that lies behind some important predictions are easy to understand.⁷

References

- [1] H. Harris. Enzyme polymorphisms in man. *Proceedings of the Royal Society of London, Series B*, 164:298–310, 1966.
- [2] J. L. Hubby and R. C. Lewontin. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics*, 54:577–594, 1966.

⁵Since the mutation rate we're talking about applies to individual nucleotide positions.

⁶That mean's they've made it more complicated.

⁷It's the concepts that get tricky, not the algebra, or at least that's what I think.

- [3] M. Kimura. Evolutionary rate at the molecular level. *Nature*, 217:624–626, 1968.
- [4] J. L King and T. L. Jukes. Non-darwinian evolution. *Science*, 164:788–798, 1969.
- [5] R. C. Lewontin and J. L. Hubby. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura*. *Genetics*, 54:595–609, 1966.
- [6] V. M. Sarich and A. C. Wilson. Immunological time scale for hominid evolution. *Science*, 158:1200–1203, 1967.
- [7] A. C. Wilson and V. M. Sarich. A molecular time scale for human evolution. *Proceedings of the National Academy of Sciences U.S.A.*, 63:1088–1093, 1969.
- [8] E. Zuckerkandl and L. Pauling. Evolutionary divergence and convergence in proteins. In V. Bryson and H. J. Vogel, editors, *Evolving Genes and Proteins*, pages 97–166. Academic Press, New York, NY, 1965.

Creative Commons License

These notes are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/3.0/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.