This Week's Citation Classic®

Jukes T H & Cantor C R. Evolution of protein molecules. (Munro H N, ed.) *Mammalian protein metabolism, III.* New York: Academic Press, 1969. p. 21-132.
[Space Sciences Laboratory, University of California, Berkeley, CA and Department of Chemistry, Columbia University, New York, NY]

This article contains an equation for use in comparing nucleotide sequences in molecules of DNA and RNA. The equation corrects for revertants, substitutions, and "multiple hits" at the same site. [The *SCI*® and *SSCI*® indicate that this chapter has been cited in over 205 publications.]

How Many Nucleotide Substitutions *Actually* Took Place?

Thomas H. Jukes
Department of Biophysics and Medical Physics
University of California
Berkeley, CA 94720

February 16, 1990

In 1965 I met Charles R. Cantor, who was 23 years old and was a graduate student in chemistry at the University of California. We started talking about molecular evolution and about comparing the polypeptide chains of homologous proteins. Charles said that a computer program should be written for searching for evidence of this, but that he was frightfully busy working on his PhD thesis, and he had no time for this. The next day he had written the program, and in 1966 we wrote two notes on its use. We resolved to write a textbook on molecular evolution together, and Charles left for Columbia University as an assistant professor in 1966. I received a request from Hamish N. Munro for a chapter in his forthcoming volume III of *Mammalian Protein Metabolism*. He asked us to write on "Evolution of protein molecules," and we sent him the manuscript that was to have been incorporated in our book. It was published in Munro's book in 1969, and the article has 110 printed pages. Citations to our long article relate only to the following short passage in it, written by Charles.

> It can be shown that the mean number of base differences at a single position on the mRNA, μ, is related to the observed fraction of residues with single base differences, p, by the expression
>
> $$\mu = \frac{3}{4} \ln \frac{3}{3-4p} \quad (1)$$

The equation (1) assumes that all single base changes (nucleotide substitutions) are equally probable and that the frequencies of all four bases in DNA are the same. This gives me the chance to point out that (1) should be called the Cantor equation, not Jukes and Cantor. The formula came into wide use when rapid DNA and RNA sequencing became available. From then on molecular biologists became interested in comparing sequences of homologous genes to study evolution. For example, a portion of the two sequences of human α and β hemoglobin genes is

α gene A C C A A C G T C A A G G C C G
C C T G G G G T A A G G T T

β gene T C T G C C G T T A C T G C C C
T G T G G G G G A A G G T G

showing 12 nucleotide substitutions (40 percent). The mean number of substitutions that has actually occurred is greater than 12, because of revertants, such as A to C to A, and multiple changes, such as A to C to G. Equation (1) corrects for these, and the probable total number of substitutions is 17 (57 percent), *not* 40 percent.

The two genes diverged from a common ancestor at least 4×10^8 years ago. Sharks go back in the fossil record for 400 million years and sharks have α and β hemoglobins (but lampreys do not). The equation tells us that the average rate of substitution per year per nucleotide site is about $0.57 \div (4 \times 10^8) = 1.4 \times 10^{-9}$. We carry with us in every red blood cell the evidence that we are in a line of descent from an ancestor who lived 400 million years ago!

An example of the use of equation (1) is in the article by C.L. Manske and D.J. Chapman.[1] These authors used the equation to correct their comparisons of 5S ribosomal RNA sequences for rev[illegible] and convergent mutations. See also references 2 and 3 for similar usage.

Charles returned to Berkeley in 1989 to direct the human genome project at the Lawrence Berkeley Laboratory.

1. **Manske C L & Chapman D J.** Nonuniformity of nucleotide substitution rates in molecular evolution: computer simulation and analysis of 5S ribosomal RNA sequences. *J. Mol. Evol.* 26:226-51, 1987. (Cited 5 times.)
2. **Aarts H J M, den Dunnen J T, Leunissen J, Lubsen N H & Schoenmakers J G G.** The γ-crystallin gene families: sequence and evolutionary patterns. *J. Mol. Evol.* 27:163-72, 1988.
3. **Tateno Y & Tajima F.** Statistical properties of molecular tree construction methods under the neutral mutation model. *J. Mol. Evol.* 23:354-61, 1986. (Cited 5 times.)

14-21