Correcting Parsimonious Trees for Unseen Nucleotide Substitutions: The Effect of Dense Branching as Exemplified by Ribonuclease[1]

Walter M. Fitch and Jaap J. Beintema†*

*Department of Ecology and Evolutionary Biology, University of California, Irvine; and †Biochemisch Laboratorium, Rijksuniversiteit Groningen

In a study of mammalian ribonuclease evolutionary rates, we applied the Fitch-Bruschi correction to reduce the bias caused by an unequal sampling of taxa in different lineages. The correction was clearly appropriate but only up to a point. The analysis showed that the sampling of taxa within the pecora was sufficiently intense that no correction for unseen, amino acid–changing, nucleotide substitutions was required. It was also found that the ribonuclease gene was duplicated at least twice at the origin of the pecoran branch of the artiodactyls.

Introduction

It has long been appreciated that parsimony procedures undercount the number of events that led to divergence of homologous gene products, and several efforts have been made to estimate the total from the number observed. The earliest was the Poisson correction of Jukes and Cantor (1969), who assumed that the substitutions were randomly distributed over the gene. This assumption is clearly violated much of the time (Fitch and Markowitz 1970; Shoemaker and Fitch 1989). For some questions it is not necessary to know the total number of substitutions, only to know that there is no methodological bias in their distribution among the branches of the tree. Molecular-clock/uniform-rate questions fall into this category. In this vein, Goodman et al. (1974) adopted a heuristic technique that used the observed excess of phyletic (tree) distance over pairwise differences in the closer parts of the tree to minimally increase the parsimonious length of longer branches. Langley and Fitch (1974) provided a second method. Most recently, Fitch and Bruschi (1987) provided a very simple method of correcting the observed change, based on the most parsimonious tree's ability to provide information on the number of new substitutions observed for every new node added to the tree. The method does not claim to estimate the total well but does claim to reduce the inequality of rates on different branches as a result of a nonrandom sampling of divergent lineages. While the method is simple, it makes the assumption that one has not already observed all of the substitutions. In the present paper we present a case where that assumption almost certainly fails. The occurrence of that failure brings with it a method of detecting that failure and a simple way of adjusting the procedure to provide the answers sought.

1. Key words: ribonuclease, unseen substitutions, mammals, parsimonious trees, gene duplication.

Address for correspondence and reprints: Walter M. Fitch, Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, California 92717.

Mol. Biol. Evol. 7(5):438–443. 1990.

0737-4038/90/0705-0004$02.00

Data and Methods

Forty-one amino acid sequences of ribonuclease from 38 species of mammals were used. All but two are pancreatic, the other two being bovine brain and semen. The sequences of 39 of them in the alignment used here were published by Beintema et al. (1986). The new sequences are those of the ox brain (Watanabe et al. 1988) and the spalax (Schüller et al. 1989) and are readily aligned to the others.

The methods employed to obtain the phylogenetic tree were those of Fitch (1971) and Fitch and Farris (1974). The relative-rate correction was according to the method of Fitch and Bruschi (1987).

Results

The analysis, using all the procedures including an alteration of the Fitch and Bruschi correction, gives the final tree shown in figure 1. The new sequences induce a few changes from the most recent previous phylogenetic study of ribonucleases (Beintema et al. 1986). Previously, joining casiragua first with coypu and then with chinchilla was equally parsimonious with joining it first with chinchilla and then with coypu, and so the former tree was depicted because that branching pattern more closely corresponds to the usual classification. In the present study the latter was more parsimonious. In the work of Beintema et al. (1986), the reindeer was grouped with the other deer to agree with the usual classification—but at a cost of an extra nucleotide substitution. Here it is shown as it would have been had the most parsimonious tree been depicted. The spalax sequence, which is new, becomes the sister taxon of the other myomorphs, which is conventional. The new paralogous bovine brain sequence joins to the bovine seminal sequence at a point indicating that the ribonuclease gene duplicated at least two times at approximately the origin of the pecorans.

The tree shows the sequence of the goat ribonuclease to have descended unchanged from the form it had in the ancestor of the bovids. That ancestor is an unresolved trichotomy. The tree also shows the topi ribonuclease to have descended unchanged from the form it had in the ancestor of the bottom nine taxa. Moreover, that ancestor is an unresolved pentachotomy. Thus, depending on how these polychotomies are resolved, there are possibly $3 \times 105 = 315$ trees of the same length as this one, 554 nucleotide substitutions. As there are $>10^{58}$ unrooted, tip-labeled, strictly bifurcating trees with 41 tips, 315 should not, perhaps, be considered to be exceptionally poor resolution.

Some of the branches have a length that is the sum of two numbers, the second of which is the Fitch-Bruschi correction. It can readily be seen that the correction has done its job in that, except for myomorphs, all the taxa are between 63 (horse and capybara) and 76 (cuis, whale, and ox brain) nucleotide substitutions from the penultimate ancestral node where the marsupial kangaroo outgroup roots the placental mammals. In the absence of the correction, again excluding the myomorphs (murids frequently seem to be evolving faster than other mammals; see, e.g., Wu and Li 1985), the distances from the penultimate node range from 19 (sloth) to 76 (ox brain). Thus the upper value is no longer 400% larger than the lower but only 21% larger.

The manner in which the Fitch-Bruschi correction is estimated is shown in the plot of the uncorrected distances from the penultimate node versus the number of intervening nodes (fig. 2). The slope of the line shows that in this tree, for each additional node up to nine, there is an increase of six in the number of substitutions.

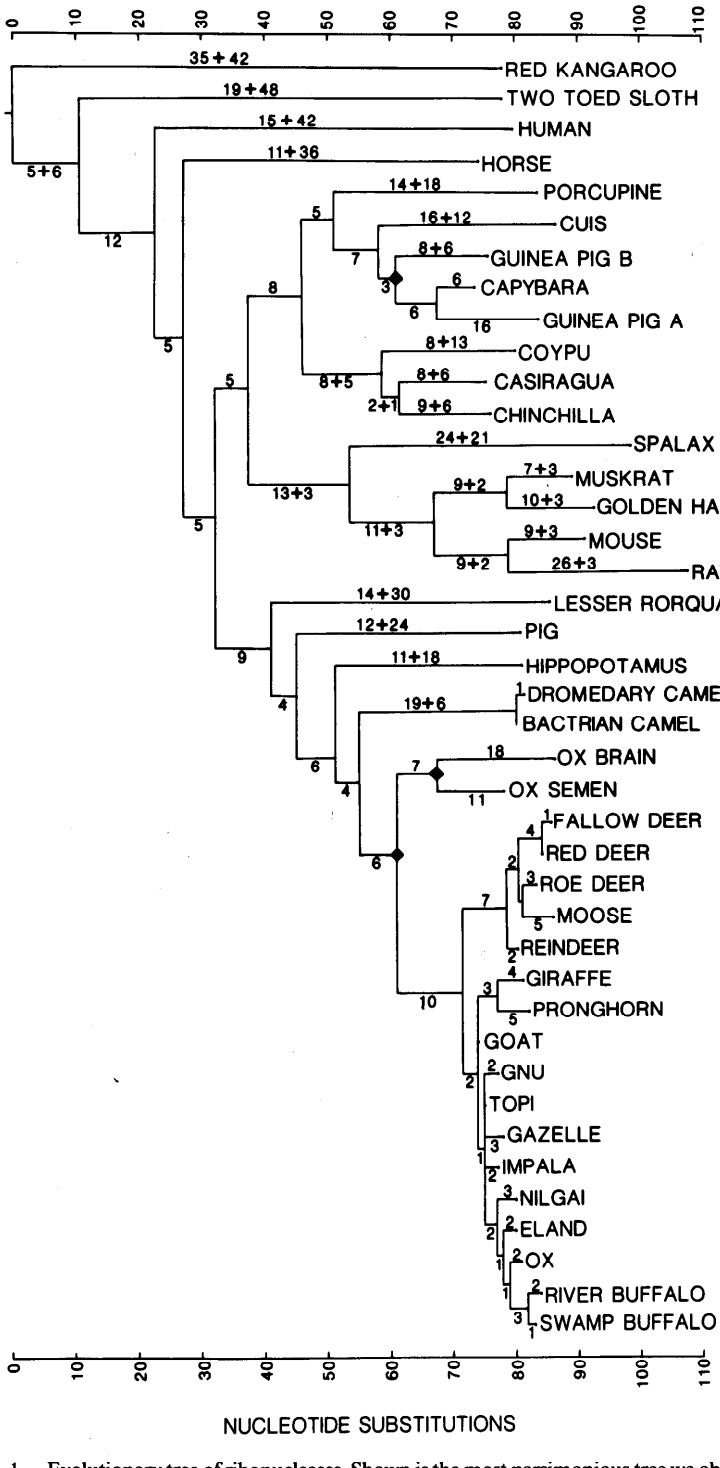

FIG. 1.—Evolutionary tree of ribonucleases. Shown is the most parsimonious tree we observed. Horizontal lengths are proportional to the corrected number of nucleotide substitutions which are shown, rounded to the nearest integer, in the form $a + b$, where a is the number of uncorrected substitutions demanded by parsimony and where b is the number of additional substitutions required by the correction procedure of

This determined the correction applied. Because the slope ceases thereafter, it was not applied to those sequences with nine or more intervening nodes.

Discussion

As a result of the correction for unequal sampling of taxa among lineages, the branch tips are more nearly of the same distance from the root, as should be the case. Thus one is not likely to conclude that rates are unequal merely as a consequence of sampling a different number of taxa in the two lineages being compared. Consequently, significant differences that remain should be more believable.

The uncorrected distances show that, since the common ancestor of mouse and rat, the 35 amino acid–changing nucleotide substitutions are divided 9 and 26 between the two lineages; the χ^2 value is minimized if we set the expected division at 17.5 substitutions each. Nevertheless, the χ^2 value is still 8.2, which, for 1 degree of freedom, has $P < 0.005$. If the corrected substitutions are used, the split is 12 and 29, with an expected division at 20.5, $X_1^2 = 7.0$, and $P < 0.009$—a still significant difference in rate in this gene in the two lines.

Another interesting rate difference is between the myomorph (e.g., mouse) and the hystricomorph (e.g., guinea pig and chinchilla) rodents. All five myomorph tips are farther from the root than are any of eight hystricomorph tips. The probability, if these were independent events, would be $2(5!)8!/13! = 0.0016$. Unfortunately, the events are not independent, and it is not clear how much of this low value is attributable to that dependency.

Figure 2 shows that the usual increasing distance between organisms as the number of intervening nodes increases comes to an abrupt halt after nine intervening nodes. This has not been observed before. Indeed, for the region of the graph from 9 to 16 intervening nodes, there is virtually no significant slope to the best-fitting straight line. This is what one would expect if all the nucleotide substitutions (those that changed the amino acids; remember these were amino acid sequences) were being caught by the parsimony procedure. In any event, the slope of zero demands that no correction be made for those points—and none was.

Is there evidence that substantially all changes among the pecorans are being detected? We believe so. The detected changes in any one branch are so few that the probability that a second change had occurred in the same position as the first change is very small. The average number of changes between nodes is only 2.3 substitutions among these 17 pecorans (bovids, pronghorn, giraffe, and deer). The number of sites

---

Fitch and Bruschi (1987). No *b* is shown if no correction is required. Vertical distances are solely to separate lineages. The diamonds (♦) represent three demonstrated gene duplications. The total number of substitutions in the tree is 554. The taxa used in the study were red kangaroo (*Macropus rufus*); two-toed sloth (*Bradypus infuscatus*); human (*Homo sapiens*); horse (*Equus caballus*); porcupine (*Hystrix cristata*); cuis (*Galea musteloides*); guinea pig (*Cavia porcellus*); capybara (*Hydrochoerus hydrochoeris*); coypu (*Myocastor coypus*); casiragua (*Proechimys guairae*); chinchilla, (*Chinchilla brevicaudata*); spalax (*Spalax ehrenhergi*); muskrat (*Ondatra zibethica*); golden hamster (*Misocricetus auratus*); mouse (*Mus musculus*); rat (*Rattus norvegicus*); lesser rorqual (*Balaenoptera acutorostrata*); pig (*Sus scrofa*); hippopotamus (*Hippopotamus amphibius*); dromedary camel (*Camelus dromedarius*); bactrian camel (*Camelus bactrianus*); ox (*Bos taurus*); fallow deer (*Dama dama*); red deer (*Cervus elaphus*); roe deer (*Capreolus capreolus*); moose (*Alces alces*); reindeer (*Rangifer tarandus*); giraffe (*Giraffa camelopardalis*); pronghorn (*Antilocapra americanus*); goat (*Capra hircus*); gnu (*Connochaetes taurinus*); topi (*Damaliscus korrigum*); gazelle (*Gazella thomsoni*); impala (*Aepyceros melampus*); nilgai (*Boselaphus tragocamelus*); eland (*Taurotragus oryx*); buffalo (swamp and river; *Bubalus bubalis*).

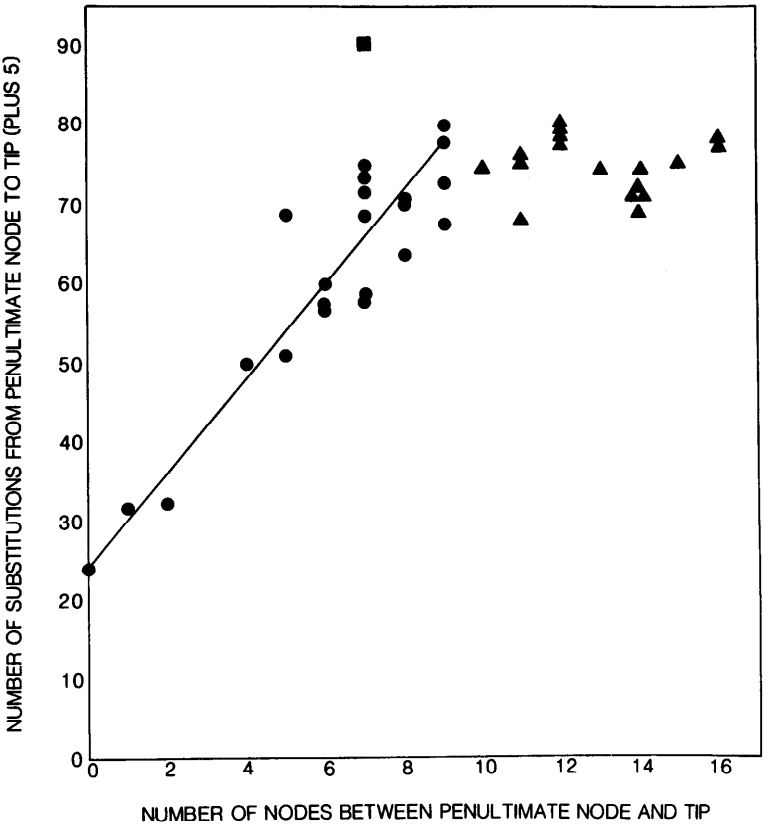

FIG. 2.—Substitutions per additional node. The plot is of the uncorrected distance (plus 5) of each taxon from the penultimate node in fig. 1 (except for the outgroup red kangaroo) against the number of intervening nodes between the penultimate node and the tips. The line is a fit to the circles. A square (■) denotes the rat sequence; the triangles (▲) denote the bottom 17 sequences of fig. 1 (fallow deer to swamp buffalo). The slope of the line is 6 substitutions/node and led to correcting distances in fig. 1 by 6 for every node <9 between the penultimate node and the tip, according to the method of Fitch and Bruschi (1987).

at which a nucleotide substitution might cause an amino acid change is $\sim 124 \times 2.5 = 310$. Not all of those sites will tolerate a change, but even if the variable positions are only 75% of that number, the frequency of substitution is still only 0.01 substitutions/variable site, so that one would expect multiple substitutions at these variable sites to be negligible. Fitch (1980) estimated the covarion number of RNAse as 77. If most of the covarion positions remained variable during the pecoran evolution, a similar rate estimate is obtained. If there were turnover among the covarions, an even smaller number of covarions would give this low rate and, hence, negligible numbers of substitutions.

Thus the Fitch-Bruschi correction procedure, although not necessarily applicable to all data sets, appears to be somewhat more general than might have been thought—in that problems of the branching being so dense that the correction should not apply are ameliorated in that those conditions are apparent, at least if they occur in sufficient numbers and in a group. A single short branch is not readily recognized as being equivalent to reducing the number of intervening nodes by one. A study of how to make such corrections is in progress, but it is a subtle problem in that the procedure,

in making use of the plot of nodes versus substitutions, is saying that this relationship would not be linear if the clock were not at least crudely uniform over the taxa studied. Once one tries for secondary correction, one is in danger of answering a different question, viz., "How many substitutions should be added to each branch to make the clock perfect?"—not "How many substitutions have we missed?" To have corrected the pecorans by the slope in the others would have made them appear to have evolved faster than the others when they had not. Going to the other extreme would destroy the unequal rates between mice and rats.

Acknowledgment

This work was supported by NSF grant BSR 8796183 to W.M.F.

LITERATURE CITED

BEINTEMA, J. J., W. M. FITCH, and A. CARSANA. 1986. Molecular evolution of pancreatic-type ribonucleases. Mol. Biol. Evol. **3**:262–275.

FITCH, W. M. 1971. Toward defining the course of evolution: minimum change for a specific tree topology. Syst. Zool. **20**:406–416.

———. 1980. Estimating the total number of nucleotide substitutions since the common ancestor of a pair of homologous genes: comparison of several methods and three beta hemoglobin messenger RNA's. J. Mol. Evol. **16**:153–209.

FITCH, W. M., and M. BRUSCHI. 1987. The evolution of prokaryotic ferredoxins—with a general method correcting for unobserved substitutions in less branched lineages. Mol. Biol. Evol. **4**:381–394.

FITCH, W. M., and J. S. FARRIS. 1974. Evolutionary trees with minimum nucleotide replacements from amino acid sequences. J. Mol. Evol. **3**:263–278.

FITCH, W. M., and E. MARKOWITZ. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixations of mutations in evolution. Biochem. Genet. **4**:579–593.

GOODMAN, M., G. W. MOORE, J. BARNABAS, and G. MATSUDA. 1974. The phylogeny of human hemoglobin genes investigated by the maximum parsimony method. J. Mol. Evol. **3**:263–278.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–126 *in* H. N. MUNRO, ed. Mammalian protein metabolism. Vol. **3**. Academic Press, New York.

LANGLEY, C. H., and W. M. FITCH. 1974. An examination of the constancy of the rate of molecular evolution. J. Mol. Evol. **3**:161–177.

SCHÜLLER, C., B. NEUTEBOOM, G. H. WÜBBELS, J. J. BEINTEMA, and A. NEVO. 1989. The amino acid sequence of pancreatic ribonuclease from the mole rat, *Spalax ehrenbergi,* chromosomal species 2n=60. Biol. Chem. Hoppe-Seyler **370**:583–589.

SHOEMAKER, J. S., and W. M. FITCH. 1989. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol. Biol. Evol. **6**:270–289.

WATANABE, H., H. KATOH, M. ISHI, Y. KOMODA, A. SANDA, Y. TAKIZAWA, K. OHGI, and M. IRIE. 1988. Primary structure of ribonuclease from bovine brain. J. Biochem (Tokyo) **104**: 939–945.

WU, C.-I., and W.-H. LI. 1985. Evidence for higher rates of nucleotide substitution in rodents than in men. Proc. Natl. Acad. Sci. USA **82**:1741–1745.

WESLEY M. BROWN, reviewing editor

Received August 31, 1989; revision received April 23, 1990

Accepted April 24, 1990