Doolittle et al. calibrated the curve, there is little difference among the curves. For levels of difference that reflect the time of the prokaryote-eukaryote split, the curves diverge significantly. Using the estimate derived from parsimony analysis without invariant sites (α = 1.21) and 70% difference, we found that an uncorrected formula underestimates the amount of divergence by 45%. Using the estimate of α = 3.29 and 18% invariant sites, an uncorrected formula underestimates divergence by almost 60%. Using the estimate of α = 0.7 derived from maximum-likelihood analysis, an uncorrected formula underestimates the number of substitutions by more than 66%. Correcting the data by Doolittle et al. present according to these estimates would date the prokaryote-eukaryote split at between 3.5 and 6 Ga. Even if we accept the contention by Doolittle et al. that at most 5% of sites are invariant, and if we use the largest estimate of α we obtained (the most generous combination of parameters possible), the date of the prokaryote-eukaryote split would be no more recent than 2.8 Ga.

J. Peter Gogarten
Lorraine Olendzenski
Elena Hilario
Department of Molecular and Cell Biology,
University of Connecticut,
Storrs, CT 06269-3044, USA

Chris Simon
Kent E. Holsinger
Department of Ecology and Evolutionary Biology,
University of Connecticut

REFERENCES AND NOTES

2. We added TPI sequences from Thermotoga maritima, Mycoplasma ficioculare, and a Vibrio species. For further description, see (5).
5. Sequences were from R. F. Doolittle et al. (1) and downloaded from the NCBI (entreze@ncbl.nih.gov) using NENTREZ and aligned using clustalw (J. D. Thompson, D. G. Higgins, T. J. Gibson, Nucleic Acids Res. 22, 4673 (1994)). Phylogenies were calculated using the distance matrix analysis implemented in clustalw (positions with gaps) and the programs PROTPARS, PRODIST, and FITCH from PHYLIP (J. Felsenstein, Phylogeny Inference Package 3.5c (1993), Univ. of Washington, Seattle, WA). In the latter two cases, gaps were encoded as missing data.
6. In none of the phylogenies calculated for dhidrofolate reductase, dihydropteroic acid dehydrogenase, phosphoglycerate kinase, and enolase did the root representatives of the archaea form the deepest (that is, closest to the eukaryotes) branch within the prokaryotes. Only in the cases of argininosuccinate synthase and porphobilogen synthase did the enlarged data sets still support the topology given by Doolittle et al. in figure 3 in their article (1).

7. We used two methods to arrive at these estimates: (i) the parsimony-based method of Gogarten et al. (Mol. Biol. Evol. 12, 988 (1995)) used for analysis of nucleotide sequences, and (ii) a maximum-likelihood method developed by Z. Yang [Phylogenetic analysis by maximum likelihood (PAAML), Version 1.1. (1996) Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park, PA]. The former method (i) does not employ a particular model of amino acid substitution, and, as expected, estimates of α derived from it (0.98 to 1.21) are substantially higher than those derived from the maximum-likelihood method (0.57 to 0.77). Maximum-likelihood estimates derived from a Dayhoff amino acid substitution model differed from those derived from a Jones model (D. T. Jones, W. R. Taylor, J. M. Thornton, Comput. Appl. Biosci. 8, 275 (1992)] by 4% or less for adenine thymophosphate (0.68 and 0.63) and hisp (0.70 and 0.76). The estimate of α from the Dayhoff model was 0.66, but only 0.57 for the Jones model with TPI. We accept α = 0.7 as a reasonable estimate for the effect of among-site rate variation from the maximum-likelihood analysis.

Response: Hasagawa and Fitch and Gogarten et al. raise the question of whether our calculated values for the most distantly related sequences (especially those involved in measuring the divergence of prokaryotes and eukaryotes) were underestimated because we did not take account of differential rates of change in individual amino acid positions ("sites"). Both comments suggest that a gamma distribution could have been employed to estimate the magnitude of a correction needed over and beyond a simple Poisson expectation. The method we used (1), however, is not an uncorrected Poisson, but employs a specified amino acid substitution table and has virtues not unlike the empirical process introduced by Dayhoff (2) that correct the other weakness in a simple Poisson: namely, the assumption that all amino acid interchanges are equally likely. The results of this correction are not unlike those obtained from the use of a gamma distribution calculated with reasonable parameters (3).

The point can best be made by comparing the relationship between distance (d) and fraction of residues changed (p) for a number of procedures, including the plots provided by Gogarten et al. which uses a gamma distribution. As pointed out in both comments, results obtained from a gamma distribution are sensitive to the choice of the α parameter and to the fraction of irre-
et al. use a low value of $a$ in one case and a high fraction of irreplaceable residues in the other to emphasize how different can be the observed and actual substitutions in the cases of a gamma distribution compared with a simple uncorrected Poisson. In contrast, attempts by others to make general corrections for a gamma distribution (3) have resulted in much more modest corrections (Fig. 1A). In this regard, Ota and Nei (3) used $a = 2$, a value which resulted in curves that closely overlay an empirical curve generated by Dayhoff et al. (2), even though the former corrects for "sites" and the latter for constraints on substitution. The value of $a = 2$ was also in close correspondence to an empirical finding in the initial study by Uzzell and Corbin (4), which suggested that a negative binomial distribution with values drawn from a gamma distribution has a better fit to rates of change in cytochrome c than does a simple Poisson.

Plots of distances calculated by our method (5) have a unique quality: namely, the distances in the nearer range are smaller than expected relative to a simple Poisson (Fig. 1B). The reason is that the most probable ("early") changes in a protein sequence tend to be of the sort that yield high substitution scores. Consider, for example, a sequence of 100 residues in which ten changes occur and for which $p$ will be close to 0.1. The $S$ values, from which the distances will be calculated, will only change modestly, and the calculated $d$ will tend to have a value near 0.05. To illustrate with the PAM-250 scoring table, a change of a valine to an isoleucine has virtually no impact on the score. As sequences become increasingly different, of course, this tendency for inconsequential change is eroded. As a consequence, the more distant sequences exhibit lower $S$ values and higher distances (Fig. 1B).

Until recently, we would have dismissed Gogarten et al.'s choice of parameters for a gamma distribution as overzealous in the extreme. However, during the course of preparing this response, we plotted the several versions of Poisson corrections recently described by Grishin (6). Unexpectedly, his theoretical formulation correcting both for rate variation at sites and for the nature of amino acid replacement did remarkably well, especially over the recent long divergence points, and the simple Poisson (Fig. 1C). The Grishin equation uses an exponential distribution of rates for different sites that is equivalent to a gamma distribution with an $a$ factor $= 1$.

What impact do all these corrections have on distance-time considerations? If $d$ values are read off the various curves (Fig. 1) at specified values of $p$, relative divergence times can be obtained directly. For example, $p = 0.31$ corresponds to the deepest divergence point (echinoderms-chordates) and the third point, and $p = 0.63$ to the plant-animal divergence (P/A) and the third point. These $p$ values stand in stark contrast to the exaggerated times suggested by Gogarten et al. in their comment. Their results may be a result, at least in part, to their assuming that their formulation is a simple Poisson, on the one hand, and, perhaps, because they did not calibrate their own curve to known fossil record divergences.

Beyond that, our method does as well as the Ota and Nei procedure (3)—which uses a gamma distribution, but which does not take account of individual amino acid variation—and also as well, or better, as the Dayhoff procedure (2), which takes account of amino acid substitution preferences, but

---

**Table 1.** Distances at some key $P$ values determined by different methods.*

<table>
<thead>
<tr>
<th>Method</th>
<th>$C/E$</th>
<th>$P/A$</th>
<th>$K/B$</th>
<th>Ratio $d$</th>
<th>Time (My)$^+$</th>
<th>Ratio $d$</th>
<th>Time (My)$^+$</th>
<th>Ratio $d$</th>
<th>Time (My)$^+$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grishin (6)</td>
<td>0.51</td>
<td>1.00</td>
<td>2.35</td>
<td>4.61</td>
<td>2536</td>
<td>2.35</td>
<td>2350</td>
<td>1.96</td>
<td>1078</td>
</tr>
<tr>
<td>Gogarten et al.</td>
<td>0.49</td>
<td>0.94</td>
<td>2.20</td>
<td>4.49</td>
<td>2470</td>
<td>2.34</td>
<td>2340</td>
<td>1.92</td>
<td>1056</td>
</tr>
<tr>
<td>Doolittle et al. (1)</td>
<td>0.27</td>
<td>0.48</td>
<td>0.88</td>
<td>3.26</td>
<td>1793</td>
<td>1.83</td>
<td>1830</td>
<td>1.78</td>
<td>979</td>
</tr>
<tr>
<td>Ota and Nei (3)</td>
<td>0.41</td>
<td>0.70</td>
<td>1.30</td>
<td>3.17</td>
<td>1744</td>
<td>1.86</td>
<td>1860</td>
<td>1.71</td>
<td>941</td>
</tr>
<tr>
<td>Dayhoff et al. (2)</td>
<td>0.40</td>
<td>0.67</td>
<td>1.24</td>
<td>3.10</td>
<td>1705</td>
<td>1.86</td>
<td>1850</td>
<td>1.68</td>
<td>924</td>
</tr>
<tr>
<td>Uncorrected Poisson</td>
<td>0.37</td>
<td>0.60</td>
<td>0.99</td>
<td>2.68</td>
<td>1474</td>
<td>1.66</td>
<td>1650</td>
<td>1.62</td>
<td>891</td>
</tr>
</tbody>
</table>

*Distances have been taken from the data plotted in Fig. 1. $^+$Time (My) obtained by multiplying the ratio by the (fossil record) divergence time of echinoderms (E) and chordates (C), which is taken as 550 My. $^*$Time (My) obtained by multiplying ratio by 1000 My, a proposed divergence time for plants (P) and animals (A).
does not take into account the rate differences at different sites. The point is best illustrated when our distances are multiplied by an empirical constant (1.5) that makes the early values \( p < 0.3 \) coincident with a simple Poisson (Fig. 1D). The range from \( p = 0 \) to \( p = 0.31 \) corresponds to the fossil record divergence times that were used to calibrate our time scale.

Because our method does not provide a correction for site variation [a limitation we acknowledged (1)], we applied a 10 to 15% correction based on a simulation experiment. Now the theoretical equation of Grishin (6) suggests that a more realistic correction would have been 25 to 30%. In no case, however—not even the “worst case” parameter selection of Gogarten et al.—do these corrections lead to prokaryote-eukaryote divergence times greater than 2500 Ma (Table 1).

On a slightly different tack, one of the virtues of restricting our study to enzymes has been that the sequences are relatively slow changing, and the overall pairwise similarities remain in a range that has usually been thought to need minimal correction (7). This property served as the fulcrum for an internal measure that we felt corroborated our general interpretation; the determination of longest distances using the 27 fastest-changing enzyme sequences gave values that were not radically different from those found with the 27 slowest changing. Because deviations are a function of dissimilarity (Fig. 1), the 27 slowest changing proteins should have given the more reliable divergence time; in fact, it put the divergence nearer to the present.

Turning to the matter of possible horizontal transfers resulting from sequences being imported during the acquisition of organelles or otherwise, in our article (1) we noted that, inadvertently, some of the sequences used may have had such a history, but unless their numbers were excessive, the results would not be greatly affected. We also implied that some of the sequences used may have been paralogs and not orthologs, noting particularly in our comparison of the slowest changing half of the set with the fastest changing half (as judged by prokaryote-eukaryote differences) that the former would be more likely to have some horizontal imports among them and the latter more likely to have paralogs. Newly available sequences make it clear that at least three of the 57 enzymes used contained bacterial paralogs (glyceraldehyde 3-phosphate dehydrogenase, glutamine synthetase, and argininosuccinate synthase). These have now been corrected, and, not unexpectedly, the corresponding distances are somewhat smaller. These changes would tend to be offset by the inclusion of any entries that turn out to be horizontal imports, which would have artifactually smaller distances reflecting a short circuit.

The relation between the archaebacteria and eubacteria needs special comment. As Gogarten et al. correctly state, our original study (1) did not have sufficient archaebacterial representation to make a firm statement about their phylogenetic position (we had called attention to the fact in footnote 25 of our article). During the interval since the submission of the manuscript for that article (April 1995), many more archaebacterial sequences have become available (8), and we have been able to increase the number of enzymes that have archaebacterial representation from nine to 16. Of the 16, we now find that in ten cases the archaebacteria cluster strongly with the eubacteria. In three others the archaebacteria sequences are more similar to eukaryotes, and in the remaining three the eubacteria and eukaryote sequences are significantly more similar to each other than either is to the archaebacteria.

Given these observations, we agree, as noted in our article, that there may have been some kind of extensive horizontal transfer among the bacteria or, possibly, a chimeric fusion involving archaebacteria on the way to the eukaryotic cell. With regard to the latter, we tend to favor the suggestions of Hartman (9) and Sogin (10) that the host cell was an early diverging entity, over the notion of Gupta and Golding (11) that a Gram-negative bacterium was involved. Whatever the case, our major findings are not much affected by the omission of any sequences that may have been involved in that hypothetical acquisition.

**REFERENCE AND NOTES**

5. The distance is equal to the negative log of similarity score corrected for the nature of the amino-acid composition, according to the formula \( d = -\ln(S_{obs} - S_{rand}/S_{obs} - S_{rand}) \), where \( S \) is obtained from an amino-acid substitution table—for example, the PAM-250 or GLSUSM-62 tables. The effectiveness of this formula to relate distance with evolutionary time has been tested by simulation with a computer model that follows the divergence of mutated protein sequences under various circumstances of constraint.
8. The overall sequence count from the 57 selected enzymes has increased from 531 to 650, mostly with new sequences from microorganisms, but also with more “bridge organisms” and fossil record representatives (D.-F. Feng, G. Cho, R. F. Doolittle, in preparation).

12 June 1996; accepted 28 August 1996

---

**Make a quantum leap.**

SCIENCE On-line can help you make a quantum leap and allow you to follow the latest discoveries in your field. Just tap into the fully searchable database of SCIENCE research abstracts and news stories for current and past issues. Jump onto the Internet and discover a whole new world of SCIENCE at the new Web address...

http://www.sciencemag.org

---

**NEW URL**

http://www.sciencemag.org
Response: Dating the Cenancester of Organisms
Russell F. Doolittle, Da-Fei Feng, Simon Tsang, Glen Cho and Elizabeth Little (December 6, 1996)
Science 274 (5293), 1751-1753. [doi: 10.1126/science.274.5293.1751]

Editor's Summary

This copy is for your personal, non-commercial use only.

Article Tools  Visit the online version of this article to access the personalization and article tools:
http://science.sciencemag.org/content/274/5293/1751

Permissions  Obtain information about reproducing this article:
http://www.sciencemag.org/about/permissions.dtl