# Determining Divergence Times of the Major Kingdoms of Living Organisms with a Protein Clock

Russell F. Doolittle, *et al.*

*Science* **271**, 470 (1996);
DOI: 10.1126/science.271.5248.470

# Determining Divergence Times of the Major Kingdoms of Living Organisms with a Protein Clock

Russell F. Doolittle,* Da-Fei Feng, Simon Tsang, Glen Cho, Elizabeth Little

Amino acid sequence data from 57 different enzymes were used to determine the divergence times of the major biological groupings. Deuterostomes and protostomes split about 670 million years ago and plants, animals, and fungi last shared a common ancestor about a billion years ago. With regard to these protein sequences, plants are slightly more similar to animals than are the fungi. In contrast, phylogenetic analysis of the same sequences indicates that fungi and animals shared a common ancestor more recently than either did with plants, the greater difference resulting from the fungal lineage changing faster than the animal and plant lines over the last 965 million years. The major protist lineages have been changing at a somewhat faster rate than other eukaryotes and split off about 1230 million years ago. If the rate of change has been approximately constant, then prokaryotes and eukaryotes last shared a common ancestor about 2 billion years ago, archaebacterial sequences being measurably more similar to eukaryotic ones than are eubacterial ones.

Estimates of when two creatures last shared a common ancestor have rested mostly on suppositions based on the fossil record. Most macrofossils are restricted to the last 600 million years, however, and phyletic assignments based on microfossils are often tenuous (*1*, *2*). As a result, the divergence times of the major groupings of biological organisms—plants, fungi, animals, protists, and bacteria—have of necessity been loose estimates fitted to the time available since the presumed first appearance of cellular life, which is thought to be about 3.5 billion years ago. The branching order of the principal lineages within that time frame has been based mainly on consideration of currently shared characters.

The advent of amino acid sequence data in the late 1950s led to the concept of a "molecular clock" (*3*) by which quantitative reconstructions of historical events might be possible. Early efforts to correlate amino acid changes with histories based on the fossil record seemed promising (*4*), and numerous studies have since been conducted that have dealt with the divergence times of all sorts of creatures (*5*). Nonetheless, the divergence times of the major groupings of organisms have remained elusive. For example, amino acid sequence–based estimates of the divergence time of prokaryotes and eukaryotes have ranged from 1.3 to 2.6 billion years (*6*, *7*). Paleontologists initially placed the divergence at 1400 million years ago (Ma) on the basis of microfossils and biogeochemistry (*8*), but more recently, swayed by data from ribosomal RNA sequences (*9*), have swung to the opposite extreme and apparently accept the notion that the prokaryotic-eukaryotic split occurred 3.5 billion years ago, shortly after life itself began (*2*).

The principal challenges to molecular clocks center around the problem of unequal rates of change over long time periods and along different lineages. Protein clocks also have the complication that different proteins change at different rates as a result of different structural and functional constraints.

Nevertheless, there is a natural tendency for homologous sequences to diverge over the course of time as a result of the mutational process, whether the changes be adaptive or neutral. Although a large number of factors enter in, the aggregate process tends to be stochastic, and, with a large enough data set, anomalies should cancel and a smooth rate of change might be effected (*10*).

During a limited pilot study to see whether protein sequences could provide a reasonable chronometry of events dating back to the last common ancestor of prokaryotes and eukaryotes (*11*), our analysis of 10 proteins with representative sequences from the major groups of organisms indicated that the last common ancestor of prokaryotes and eukaryotes existed 1.9 ± 0.6 billion years ago. The uncertainty was largely a reflection of the small number of appropriate sequences available for comparison.

Now we have expanded that study to 57 different proteins comprising 531 different sequences. In addition, we have analyzed the data in several ways including tests for self-consistency among the data themselves, adjustments for observed changes in rate along different lineages, and corrections for the way in which amino acid sequences change over long periods of time. In the end, a wholly plausible set of divergence times has emerged for all the major biological kingdoms.

## Choosing the Data Set

Although there are a great many sequences in current databases, in only a relatively small number of cases is the "same" protein broadly represented. There also can be confusion about whether proteins with the same name from two different organisms are really related or are merely functionally equivalent (*12*). A further complication is unintended comparisons of paralogous rather than orthologous descendants (*13*), and even an occasional horizontal gene transfer (*14*). We have limited our study to the sequences of enzymes, partly because the nomenclature for enzymes is reasonably systematic, and partly because many enzymes occur in most organisms. Accordingly, we devised a procedure for screening all enzyme sequences in a database to see which ones were suitably represented and useful for comparative analysis (*15*–*23*). In the end, 57 enzyme groups totaling 531 sequences survived the screening process (Table 1).

The 531 amino acid sequences were from 15 principal groups of organisms, including nine animal subgroups, fungi, plants, slime mold, protists, archaebacteria, and eubacteria. The nine animal groups included six vertebrate types (placental mammal, marsupial, bird-reptile, amphibian, fish, and cyclostome), a seventh deuterostome (echinoderm), schizocoelomates (arthropods, mollusks), and pseudocoelomates (nematodes). The mammalian group was subdivided by order, and as a result we actually considered 15 divergences, beginning with the radiation of mammalian orders. Subsequently, we also subdivided the eubacteria into Gram-negative and Gram-positive organisms for a consideration of when that divergence may have occurred.

## Sequence Resemblances Among the Major Groups

The first phase of our study entailed the proper alignment of groups of enzyme sequences and determination of within-group similarities. Percent identity was used as a familiar, if rough, index of similarity; it was defined as the number of identical residues in two aligned sequences divided by the total number of matched residues.

The authors are at the Center for Molecular Genetics University of California, San Diego, La Jolla, CA 92093–0634, USA.

*To whom correspondence should be addressed.

On the average, plant sequences are more like animal sequences than are fungal ones (Fig. 1, A and B). This is true whether the entire data set is considered or only those 30 enzymes for which sequences were available from all three groups. The distribution of similarities is remarkably tight, the range of identities between animal and plants covering the span from 39 to 72 percent identity (mean, 57; SD, 8). The 54 comparisons between fungi and animal sequences ranged from 36 to 69 percent identity (mean, 55; SD, 8). Comparable results were obtained when comparisons were restricted to the subset of 30 enzymes for which representatives from all three groups were available. The average similarity of plant and fungal sequences was just about the same as the animal-fungal value (Fig. 1C). The difference between plants and animals and fungi and animals was only marginally significant (1.6 SD by the Student's *t* test), plant sequences being more similar to animal sequences in 18 of 30 comparisons.

When all 57 enzyme sets were analyzed, the 120 sequences from eubacteria and 146 from eukaryotes were found to average 37 percent identity with the full range covering a span from 20 to 56 percent identity (Fig. 1D). The results compare favorably with those of a previous study in which 28 enzyme and 2 non-enzyme sequences from *Escherichia coli* and humans were 34 percent identical, on the average (*24*). At 39 percent identity, archaebacterial sequences were more similar to those of eukaryotes than they were to those of eubacteria. The differences were apparent whether all available sequences were considered or only those nine subsets that contained both archaebacterial and eubacterial sequences, but the statistical significance was marginal (*25*).

## Calculating Distances from Sequence Resemblances

It is well established that protein sequence comparisons are more informative when weights are used that take into account structural and genetic biases for amino acid replacements. A number of amino acid substitution matrices have been generated or compiled by various means, the most popular of which has been the Dayhoff PAM-250 scale (*20*). Some other more recently introduced scales include the GCB (Gonnet-Cohen-Benner) matrix (*21*) and the BLOSUM tables (*22*). Although weighted scales have little bearing on either alignments or phylogenies when sequences are more than 30 percent identical (*26*), which is the case for most of the alignments used in this study, we still thought it prudent to try various weight matrices to ensure against some hidden bias. In all cases the similarity scores obtained were scaled as follows (*26*):

$$S = (S_{obs} - S_{rand})/(S_{ident} - S_{rand})$$

where $S_{obs}$ is the observed similarity score obtained by summing the scores for two aligned amino acids obtained from the weight matrices, $S_{rand}$ the corresponding score for two random sequences of the same lengths and compositions, and $S_{ident}$ the average score of the two self-comparisons. The scoring system corrects for chance matches and relates the course of sequence

**Table 1.** Enzyme sequences used for comparisons.

| E.C. number | Name | Length* | *N* | Animals | Plants | Fungi | Protists | Bacteria |
|---|---|---|---|---|---|---|---|---|
| 1.1.1.205 | Inosine monophosphate dehydrogenase | 337 | 10 | 4 | | 1 | 2 | 3 |
| 1.1.1.27 | L-Lactate dehydrogenase | 306 | 15 | 6 | 2 | | | 7 |
| 1.1.1.34 | HMG-CoA reductase | 403 | 15 | 6 | 5 | 1 | | 1 |
| 1.1.1.42 | Isocitrate dehydrogenase | 406 | 6 | 1 | 2 | 1 | | 2 |
| 1.1.1.49 | Glucose 6-phosphate dehydrogenase | 483 | 9 | 3 | 1 | 1 | 1 | 3 |
| 1.15.1.1 | Superoxide dismutase (Cu-Zn) | 153 | 18 | 8 | 5 | 2 | | 3 |
| 1.17.4.1 | Ribonucleotide reductase (small subunit) | 380 | 6 | 3 | | 1 | 1 | 1 |
| 1.17.4.1 | Ribonucleotide reductase (large subunit) | 751 | 6 | 3 | | 1 | 1 | 1 |
| 1.2.1.12 | Glyceraldehyde 3-phosphate dehydrogenase | 292 | 20 | 7 | 3 | 4 | 3 | 3 |
| 1.2.1.3 | Aldehyde dehydrogenase | 468 | 12 | 5 | 1 | 3 | | 3 |
| 1.2.4.1 | Pyruvate dehydrogenase | 322 | 8 | 5 | 1 | 1 | | 1 |
| 1.2.4.2 | 2-Oxoglutarate dehydrogenase | 202 | 6 | 1 | | 1 | | 4 |
| 1.3.3.1 | Dihydroorotate oxidase | 291 | 6 | 2 | 1 | 1 | | 2 |
| 1.4.4.2 | Glycine dehydrogenase (decarboxylating) | 899 | 4 | 2 | 1 | | | 1 |
| 1.5.1.3 | Dihydrofolate reductase | 160 | 17 | 5 | 2 | 2 | 5 | 2 |
| 1.8.1.4 | Dihydrolipoamide dehydrogenase | 456 | 9 | 2 | 1 | 1 | 1 | 3 |
| 2.1.1.45 | Thymidylate synthase | 286 | 13 | 2 | 2 | 2 | 4 | 3 |
| 2.1.1.63 | Cysteine S-methyl transferase | 177 | 6 | 2 | | 1 | | 3 |
| 2.1.2.1 | Glycine hydroxymethyl transferase | 457 | 10 | 2 | 2 | 2 | | 4 |
| 2.1.3.2 | Aspartate carbamoyl transferase | 309 | 10 | 2 | 2 | 1 | | 4 |
| 2.1.3.3 | Ornithine carbamoyl transferase | 324 | 9 | 3 | | 2 | | 4 |
| 2.3.1.12 | Dihydrolipoamide S-acetyl transferase | 423 | 7 | 2 | | 1 | | 4 |
| 2.3.1.16 | Acetyl CoA C-acetyl transferase | 384 | 8 | 2 | 1 | 2 | | 3 |
| 2.4.1.18 | 1,4-α-glucan branching enzyme | 605 | 8 | 1 | 2 | 1 | | 4 |
| 2.5.1.1 | Dimethylallyl transferase | 295 | 7 | 2 | 1 | 1 | | 1 |
| 2.5.1.6 | Methionine adenosyl transferase | 393 | 8 | 3 | 3 | 1 | | 1 |
| 2.6.1.1 | Aspartate transaminase | 411 | 9 | 4 | 3 | 1 | | 1 |
| 2.6.1.16 | Glutamine fructose 6-phosphate transaminase | 653 | 4 | 1 | | 1 | | 2 |
| 2.7.1.11 | Phosphofructokinase | 285 | 11 | 4 | | 2 | | 5 |
| 2.7.1.40 | Pyruvate kinase | 457 | 15 | 3 | 2 | 5 | 1 | 3 |
| 2.7.2.3 | Phosphoglycerate kinase | 407 | 14 | 5 | | 2 | 2 | 4 |
| 2.7.4.6 | Nucleoside diphosphate kinase | 149 | 9 | 2 | 3 | 1 | | 2 |
| 2.7.6.1 | Phosphoribose pyrophosphokinase | 310 | 5 | 1 | | 1 | | 3 |
| 2.7.7.6 | DNA-directed RNA polymerase | 284† | 10 | 3 | 1 | 1 | 2 | 1 |
| 3.1.3.1 | Alkaline phosphatase | 322 | 7 | 4 | | 1 | | 2 |
| 3.1.3.11 | Fructose bisphosphatase | 329 | 12 | 3 | 2 | 2 | | 3 |
| 3.2.1.22 | Alpha-galactosidase | 297 | 7 | 1 | 2 | 2 | | 2 |
| 3.4.21.4 | Trypsin | 217 | 12 | 9 | | 1 | | 2 |
| 3.6.1.23 | dUTP pyrophosphatase | 151 | 5 | 1 | 1 | 1 | | 1 |
| 4.1.1.23 | Orotidine phosphate decarboxylase | 237 | 13 | 3 | 1 | 5 | | 3 |
| 4.1.1.32 | Phosphoenolpyruvate carboxykinase | 504 | 8 | 6 | | 1 | | 1 |
| 4.1.1.37 | Uroporphyrinogen decarboxylase | 359 | 6 | 2 | | 1 | | 3 |
| 4.2.1.11 | Enolase | 366 | 18 | 8 | 3 | 3 | 1 | 2 |
| 4.2.1.24 | Porphobilinogen synthase | 322 | 9 | 2 | | 1 | | 2 |
| 4.3.2.1 | Argininosuccinate lyase | 454 | 7 | 3 | 1 | 2 | | 1 |
| 5.1.3.2 | Uridine 5′-diphosphate–glucose 4-epimerase | 340 | 7 | 1 | | 2 | | 4 |
| 5.2.1.8 | Peptidyl prolyl isomerase | 161 | 12 | 5 | 2 | 3 | | 2 |
| 5.3.1.1 | Triose phosphate isomerase | 229 | 21 | 9 | 3 | 3 | 2 | 4 |
| 5.99.1.3 | DNA topoisomerase (adenosine triphosphate–hydrolyzing) | 463 | 11 | 4 | | 2 | 2 | 3 |
| 6.1.1.3 | Threonine-tRNA ligase | 645 | 4 | 1 | | 1 | | 2 |
| 6.1.1.5 | Isoleucine-tRNA ligase | 935 | 6 | 1 | | 1 | | 3 |
| 6.1.1.9 | Valine-tRNA ligase | 463 | 5 | 1 | | 2 | | 2 |
| 6.1.1.21 | Histidine-tRNA ligase | 431 | 5 | 2 | | 1 | | 2 |
| 6.3.1.2 | Glutamate-ammonia ligase | 323 | 14 | 5 | 3 | 1 | | 5 |
| 6.3.4.4 | Adenylosuccinate synthase | 426 | 6 | 2 | | 1 | | 2 |
| 6.3.4.5 | Argininosuccinate synthase | 399 | 8 | 3 | | 1 | | 2 |
| 6.3.5.4 | Asparagine synthase | 555 | 6 | 2 | 3 | | | 1 |

*Average length used. †Only portions of sequences used.

divergence to a true first-order decay process. These scores were subsequently transformed into distance ($D$) measures by the Poisson relationship (*27–31*):

$$D = -\ln S \times 100$$

Our strategy for determining the divergence times with distance data depended on two quite different operations. In the first, the main goal was to obtain approximate times by extrapolation of a line based on the vertebrate fossil record. A constant rate of change was presumed throughout, and the possibility of different rates of change for different lineages was not considered. We also ignored the fact that not every enzyme group was represented in every biological grouping, but relied instead on the data being sufficiently abundant to fall within the realm of the Law of Large Numbers (*32*), a proposition we tested by sampling the data in various ways.

The second phase of the analysis was a refining process that took into account factors ignored in the first stage. Phylogenetic analysis was used to determine different rates of change for the various lineages, as well as to determine proper branching orders for those divergences that took place within relatively short periods of time. The impact of different enzymes tending to change at different rates was taken into account by normalizing the data in the various subsets by comparing components common to them all.

Finally, we considered the possibility that a linear relation between our calculated distances and evolutionary time might not be wholly valid. We therefore made an estimate of how different the divergence times would be if distance values were corrected for various fractional contents of irreplaceable or slowly changing residues in the proteins under study.

## Fixing Divergence Times

Even with the aid of a fossil record, there is always uncertainty in fixing a divergence time; the fossil record can only provide a "first appearance." Nevertheless, our plan was to establish a baseline rate with sequences from vertebrate animals, for which there is a reasonably good fossil record (*33*), and then to extrapolate that rate to obtain the other divergence points (Tables 2 and 3).

We initially examined slopes obtained separately by comparisons based on the PAM-250 and BLOSUM-62 matrices. The PAM-250 plot put the plant-animal-fungi junctions near a billion years ago (Fig. 2A), but the BLOSUM plot had a steeper slope and those junctions appear to be somewhat more recent (Fig. 2B). Because of the way the two weighting scales were originally designed (*20, 22*), the PAM-250 data should be more reliable for sequences that are more than 50 percent identical and the BLOSUM-62 data should be better for sequences less than 50 percent identical. Accordingly, the averaged values of the PAM and BLOSUM data were plotted with the initial PAM slope, and a set of divergence times was obtained from the observed distances (Fig. 2C). The percentages of identities were then plotted against the complete set of time points (Fig. 2D).

Simple extrapolation of the distance line led to a divergence time for the deuterostomes and protostomes of about 700 Ma (Fig. 2C). The BLOSUM comparisons indicated that the schizocoelomate (predominantly *Drosophila*) and pseudocoelomate (represented among these data mostly by *Caenorhabditis elegans* sequences) animals diverged at about the same time, but the PAM comparisons had the schizocoelomates emerging more recently. The latter result was confirmed by a thorough consideration of all intergroup distances by the subset strategy (see below). Our best estimate of the deuterostome-protostome divergence is 670 Ma, with the schizocoelomate-pseudocoelomate divergence occurring 50 to 100 Ma before that. Although these estimates are somewhat greater than most textbook values, they seem consistent with recent evaluations of the fossil record

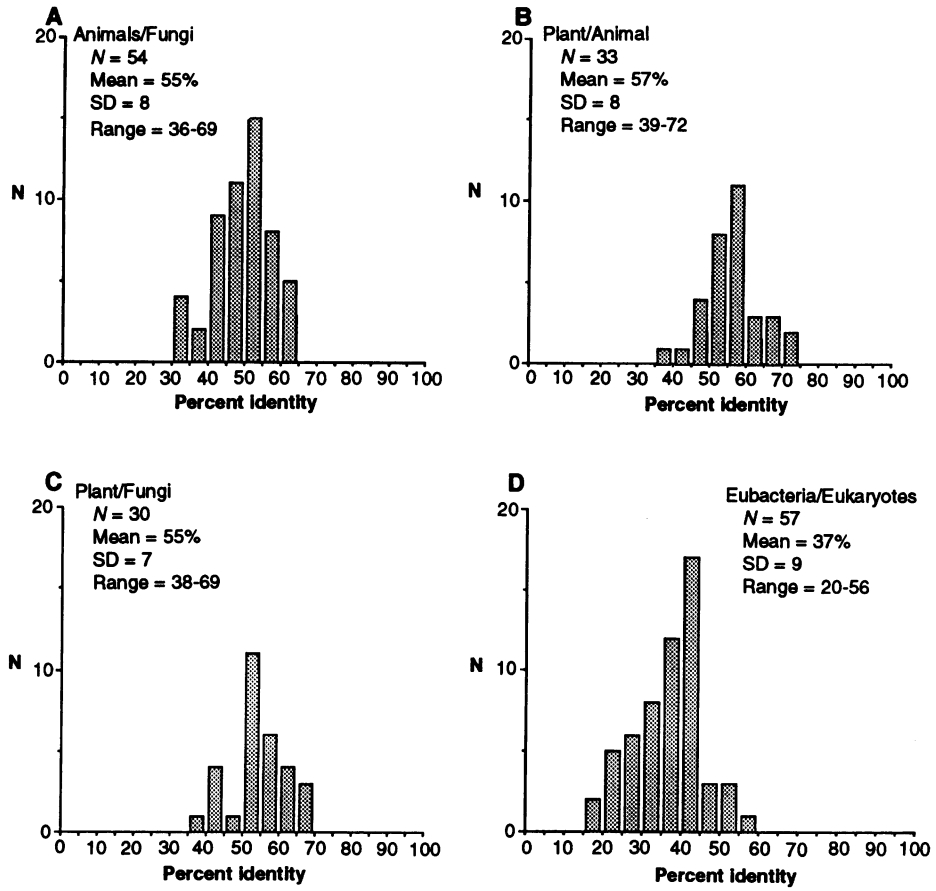


**Fig. 1.** Resemblances (percent identity) of enzyme sequences from principal biological groups as measured in blocks of five percentage points.

**Table 2.** Average resemblances and divergence times from fossil record.

| | *N** | Identity† (% ± SD) | Distance‡ | LCA§ (Ma) |
|---|---|---|---|---|
| Mammal-mammal | 43 | 91 ± 6 | 6 | 100 |
| Eutheria-marsupial | 2 | 92 ± 2 | 5 | 130 |
| Mammal-bird-reptile | 12 | 84 ± 6 | 11 | 300 |
| Amniote-amphibian | 5 | 78 ± 9 | 17 | 365 |
| Tetrapod-fish | 4 | 74 ± 8 | 22 | 400 |
| Gnathostome-lamprey | 1 | 78 | 16 | 450 |
| Chordate-echinoderm | 1 | 69 | 27 | 550 |

*Number of enzyme sets compared. †Percent identity. ‡Distances taken from Fig. 2C. §Last common ancestor.

that suggest the existence of pre-Ediacaran metazoans (*34*).

In line with their being more similar to animal sequences, plants appeared on the distance line ahead of the fungi (Fig. 2). When these data were subjected to phylogenetic analysis, however, fungi and animals clustered in every instance, no matter which subset was studied (B, D, F, G, I, or K in Table 4). Simple inspection of intergroup distances makes it evident that the sequences from fungi have been changing faster than those of plants and animals (*35*). These observations are in full accord with recent reports that suggest that animals and fungi are more recently related than animals and plants (*36*).

The 29 protist sequences used were mainly represented by kinetoplastid organisms, especially trypanosomes, leishmania, and crithidia. On average, the differences between protists and the principal kingdoms (plants, animals, and fungi) were only slightly greater than distances between members of the kingdoms (Table 3). Although the protists are likely a polyphyletic group, it is clear that the ones we used last shared a common ancestor much more recently than the divergences of eukaryotes from prokaryotes; extrapolation of the distance line puts the protist divergence at about 1230 Ma. Phylogenetic analysis of subsets C, D, and G (Table 4) revealed that the average rate of change for protist sequences has been about 35 percent greater than the rates for animal and plant lineages. As a result, the corrected divergence time appears somewhat more recent (Table 3).

Contrary to this result, the microfossil record is reported to have forms resembling protists appearing as early as 1700 Ma (*37*). However, our data set may not have a truly representative set of protists, and our estimated late divergence time may reflect that sampling bias. In this regard, three sequences from *Giardia lamblia*, frequently cited as a very early diverging eukaryote (*38*), were no more different from those of the higher eukaryote group than other protists.

Most systematicists classify the slime mold, *Dictyostelium discoideum*, as a protist (*39*) although a set of eight slime mold protein sequences was reported to be much more similar to those of higher eukaryotes than would be expected for a genuine member of that group (*40*). Our results tend to confirm those findings although the degree of confidence is limited because the number of sequences is small ($N = 5$), and equivalent sequences from other protists were not available for direct comparison. Nonetheless, the data indicate that *Dictyostelium* diverged from the main line more recently than protists and at about the same time as plants (Fig. 2C).

Direct extrapolation of the distance line indicates that eukaryotes last shared a common ancestor with archaebacteria 1800 Ma, and with eubacteria slightly more than 2000 Ma (Fig. 2C). These values are in accord with reports of microfossils whose age is 1700 to 1900 million years and that resemble eukaryotic cells (*41*), but they are at odds with the claim of a 2100-million-year-old fossil alga thought to resemble extant chloroplast-containing eukaryotes (*42*).

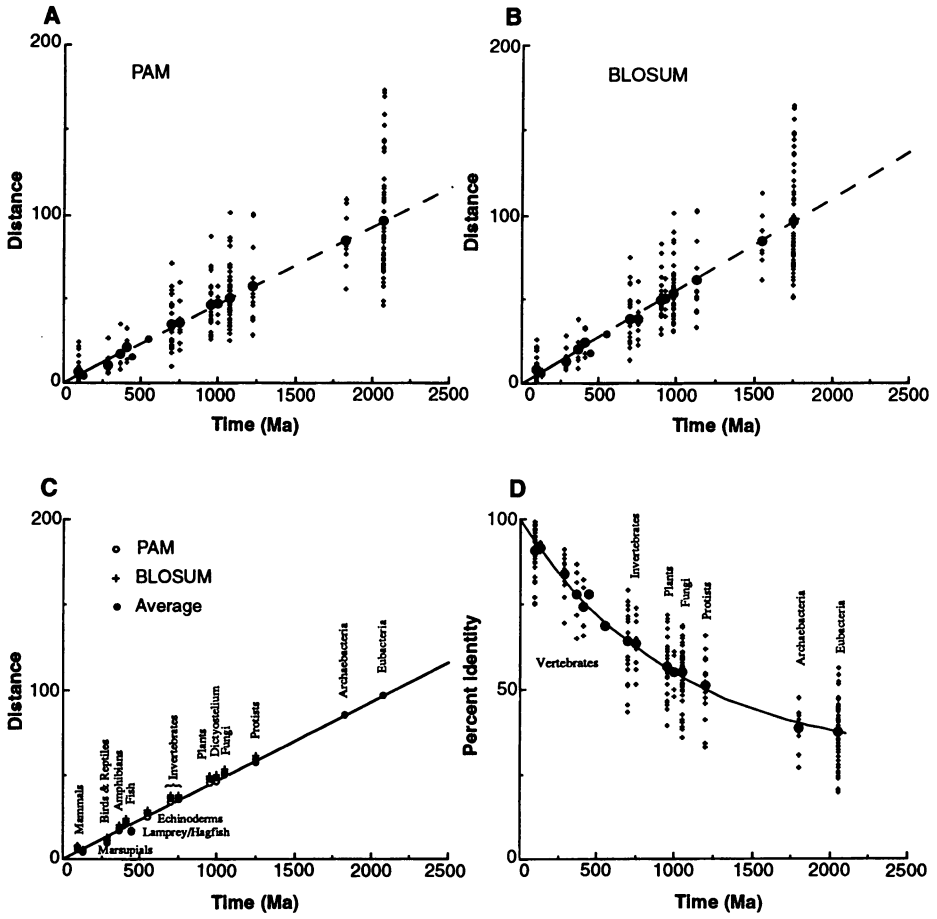


**Fig. 2.** Calculated distances determined with PAM-250 and BLOSUM-62 weighting scales plotted as a function of divergence time. (**A** and **B**) Slopes determined from the major animal divergences based on the fossil record and constrained to pass through the origin. Large symbols are the averages of all the individual data points (small symbols). Dashed lines denote extrapolations to which the distance points were fitted. (**C**) Slope based on vertebrate PAM values, but data points are averages of both PAM and BLOSUM values. Distances for each enzyme were calculated between the sequence for a given taxon and all other taxa more recently diverged from the trunk; except that sequences from plants, fungi, and *Dictyostelium* were compared only with the corresponding animal sequences (that is, at this stage no position was taken with regard to the branching order of these three groups), and, similarly, sequences from archaebacteria and eubacteria were only compared with the corresponding sequences from eukaryotes and not with each other. Because plants and slime mold gave the same distances relative to animals, they are plotted side by side. (**D**) percent identities plotted against divergence times taken from (C).

**Table 3.** Average resemblances and divergence times by extrapolation.

| | *N** | ID† (%) | D‡ | LCA§ | LCA′ | LCA″ | LCA‴ |
|---|---|---|---|---|---|---|---|
| Deuterostome-protostome | 21 | 64 ± 10 | 36 | 750 | 656 | 675 | 675 |
| Schizocoelome-pseudocoelome | 9 | 64 ± 8 | 37 | 750 | 784 | 750 | 750 |
| Fungi-animal | 54 | 55 ± 8 | 52 | 1050 | 978 | 965 | 965 |
| Plant-animal | 33 | 57 ± 8 | 47 | 1000 | 1000 | 1000 | 1000 |
| Protist–plant-animal-fungi | 14 | 51 ± 10 | 59 | 1250 | 1236 | 1230 | 1230 |
| Archaebacteria-eukaryotes | 9 | 39 ± 6 | 85 | 1800 | 1889 | 1700 | 1870 |
| Eubacteria-eukaryotes | 57 | 37 ± 9 | 96 | 2050 | 2080 | 1875 | 2156 |
| *Bacilli–E. coli* | 28 | 45 ± 9 | 75 | (1500) | 1610 | 1450 | 1523 |
| *E. coli–Salmonella* | 8 | 94 ± 6 | 6 | (100) | (100) | (100) | (100) |

*Number of enzyme sets compared. †Percent identity ± SD. ‡Distances, from Fig. 2C. §LCA, last common ancestor given as million years ago (LCA from Fig. 2C); LCA′, average of Fig. 3, A and B; LCA″, after scaling (Fig. 3C); and LCA‴, after correction for amino acid replacement constraints.

Subset analysis (below) was consistent with the archaebacteria being grouped with the eukaryotic lineage and supports other protein sequence comparisons, especially those that have taken advantage of early gene duplications, showing that at least some archaebacterial proteins are more closely related to eukaryote than to eubacteria proteins (43). Phylogenetic analysis of all the data placed the root between the archaebacteria and the eubacteria, and a negative branch length resulted when attempts were made to group the archaebacteria with the eubacteria. The data also show that the rate of change of archaebacteria sequences is similar to the eukaryote rate, as determined by the "relative rate test" (35). Furthermore, the sequences from the eubacteria also appear to be changing at about the same rate, so long as the root is placed in accordance with the extrapolated distance line.

The divergence time of Gram-positive and Gram-negative bacteria was estimated by two different comparisons: in one, 51 sequences from Gram-positive organisms were compared with 84 sequences from Gram-negative organisms (Fig. 3, A and B). The other comparison included 28 enzymes common to the genus *Bacilli* and to *E. coli*. In both comparisons, the two groups were 45 percent identical, and the calculated divergence time was about 1450 Ma (Table 3).

Those eubacteria that are not usually classified as either Gram-positive or Gram-negative were also examined. This group, which included five cyanobacteria, was no more different from the Gram-positive and Gram-negative than were the latter from each other. Apart from emphasizing that all the eubacteria represented in our study are monophyletic, the result may reflect a commonality of genomic exchange among eubacteria (*14*).

Comparison of nine enzymes common to *E. coli* and its close relative, *Salmonella typhimurium*, revealed that, at 94 percent identity, they were just slightly less similar than are the same enzymes from various mammalian orders (95 percent identical, on the average), a result in good agreement with an earlier estimate that the divergence between these bacterial groups occurred 100 to 130 Ma (*44*). We therefore conclude that the rate of sequence change per unit time among the enterobacteria is not significantly different from that observed in animals.

We cannot be certain that all the sequences analyzed in this study are truly orthologous within their group. Nor can we be certain that an occasional horizontally transferred sequence has not crept into the collection. Indeed, the enzyme with the highest resemblance between eukaryotes and eubacteria, phosphoenol pyruvate carboxykinase (E.C. 4.1.1.32), is hardly any more similar when fungi and animals are compared (no plant or protist sequences are yet available), and some kind of horizontal transfer may have occurred. But we think that the number of comparisons made was sufficiently large that such anomalies, if they exist, would have little impact. To test this proposition, we sampled the data in various ways to see what effect the omission of certain sequences would have on extrapolated divergence times. For example, we analyzed 10 data sets in which seven randomly chosen enzyme groups were omitted each time; the operation had no significant effect on the average results (Table 5). We also removed the seven fastest changing sets of sequences and, in another case, the seven slowest. The former had virtually no effect, and the latter moved the prokaryote-eukaryote junction nearer to the present (Table 5). In addition, we divided the 54 enzyme sets that contained animal, fungi, and eubacterial sequences into two groups, the 27 fastest changing and the 27 slowest. The results were only marginally affected, the more conservative proteins moving the boundary nearer to the present by less than 10 percent and the faster changing ones moving it further back in time by about the same amount (Table 4).

**Table 4.** Some subsets of common sequences.

| Subset | Biological groups* | *N*† | SF‡ |
|---|---|---|---|
| A | Animal-fungi-eubacteria | 54 | 1.00 |
| B | Animal-fungi-plant-eubacteria | 30 | 1.03 |
| C | Anima-fungi-protists-eubacteria | 14 | 1.09 |
| D | Animal-fungi-plant-protists-eubacteria | 9 | 1.24 |
| E | Animal-fungi-archaebacteria-eubacteria | 9 | 0.98 |
| F | Animal-fungi-plant-archaebacteria-eubacteria | 5 | 0.96 |
| G | Animal-fungi-plant-protists-archaebacteria-eubacteria | 4 | 1.14 |
| H | Deuterostomes-schizocoeles-fungi-eubacteria | 21 | 1.00 |
| I | Deuterostomes-schizocoeles-fungi-plant-eubacteria | 13 | 1.06 |
| J | Deuterostomes-schizocoeles-pseudocoeles-fungi-eubacteria | 7 | 1.20 |
| K | Deuterostomes-schizocoeles-pseudocoeles-fungi-plant-eubacteria | 6 | 1.21 |
| L | Animal-fungi-*Bacilli*–*E. coli* | 28 | 1.08 |

*Deuterostomes are chordates, echinoderms; schizocoeles are arthropods, annelids, and others; pseudocoelomates are nematodes, and others. †*N*, the number of enzyme types present in at least one member of each lineage in a subset. Thus, 54 of the enzymes are common to subset A, but only four enzymes are common to subset G. ‡SF, scale factors B to G and L based on fungi-eubacteria distances relative to set A. Scale factors for subsets H to K based on animal-fungi distances relative to set A.

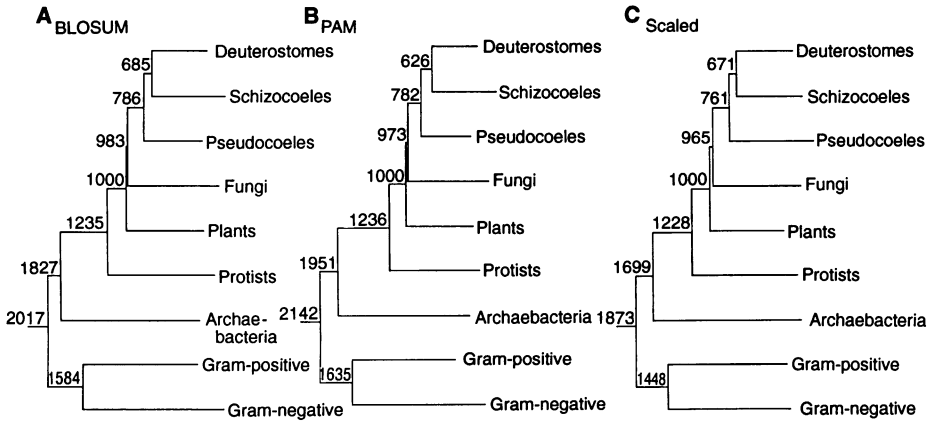


**Fig. 3.** Overall phylogenies calculated from all intergroup distances. (**A**) The phylogeny was calculated from intergroup raw data determined with the BLOSUM substitution matrix. (**B**) The tree was calculated with the equivalent raw data derived from the PAM-250 matrix. (**C**) The phylogeny was calculated from scaled data that were derived from both PAM and BLOSUM weighting and averaged. Scaling was based on subset A (animal-fungi-eubacteria), members of which occur in all the other subsets. The animal-fungi distance from subset A was used to scale all the animal intergroup distances, and the fungi-eubacteria distance from subset A to scale all other intergroup distances. The numbers at the nodes indicate divergence times in millions of years, based on the plant-animal divergence being set equal to 1000 Ma.

## Scaling and Rate Adjustments

Overall phylogenies for the principal biological groups were calculated (*45*) from the raw data derived from the BLOSUM and PAM weighting scales (Fig. 3, A and B). The calculations used all the available data (526 sequences from 57 enzymes; ~5 slime mold sequences were omitted), whether or

not every taxon was represented. Again, our justification for this application is the law of large numbers (*32*). The phylogenetic trees were surprisingly robust.

Although the ideal data set would have included a complete representation of all 15 biological groups for all 57 enzymes, such completeness in current databases is not yet at hand. Nonetheless, it was possible to assemble numerous subsets of the data that were complete unto themselves. For example, sequences were available from 30 of the enzymes for the four major kingdoms—animals, fungi, plants, and eubacteria. This set of 30 common sequences was used to determine distances between groups and to construct phylogenies, which in turn were examined in the light of the gross divergence times measured by aggregate averages and vice versa. Other smaller subsets (Table 3) were treated similarly. Relative rates determined by subset analysis were used to correct the aggregate data and adjustments in branching order were made if needed. Scaling factors (SF) were determined by normalizing the inter-pair distances for the three taxa (animals, fungi, and eubacteria) that were common to all subsets. In this way, it was possible to construct a corrected phylogeny for all groups with consistent divergence times assigned to each node.

A corrected phylogeny was then calculated with the scaled distances determined by the subset strategy, whereby distances between groups from various subsets were scaled and averaged, and an overall phylogeny computed that yielded a self-consistent set of divergence times (Fig. 3C). The most obvious difference realized by scaling was apparent in the lineage leading to present-day pseudocoelomates (for example, *C. elegans*), and here caution must be extended in that the scaling was derived from relatively small subsets (subset J has only seven members, and subset K, which is a subset of J, has only six). Beyond that, scaling had only a modest impact on the relative branch lengths. Nonetheless, the scaled values are the more rigorously determined and were used for the final assignment of divergence times (Fig. 3C and Table 3). In general, the adjustments tended to move the older divergences nearer to the present, the natural consequence of several lineages changing faster than the sequences from animals used to calibrate the distance line. Similarly, the junctions of eukaryotes with archaebacteria and eubacteria were moved forward in time by about 10 percent after all adjustments were incorporated (Fig. 3C).

## Time and Distance Considerations

Even with scaling and relative rate corrections, these divergence times depend on a linear correspondence between the distances calculated from sequence similarities and absolute time. As noted above, the Poisson condition is based on the assumption that the likelihood of replacement is the same for all residue positions, something we know is not true. Even the most changeable of amino acid positions can have constraints (*4*). The question is whether the effect of differential replacement is significant, an issue often debated (*30, 47*). Most enzymes have essential residues that cannot be replaced under any circumstances without loss of function. However, the number of such residues is usually small relative to the numbers of residues that can be changed more freely, and there are enzymes where homology has been confirmed only on the basis of three-dimensional structures, virtually all sequence resemblance having been eroded (*48*).

Nonetheless, it is a simple matter to correct the Poisson relation for various fractions of irreplaceable residues (*49*), and we reconsidered the extrapolated data in this light. Thus, if the irreplaceable fraction were a reasonable 0.05 to 0.10, our data still fall within the realm of a linear extrapolation. Even as large a fraction as 0.15 would extend the divergence time for eukaryotes and eubacteria by only 10 percent, barely offsetting the corrections imposed for variable rates of change and scaling. In contrast, if the eukaryote-prokaryote divergence occurred 3500 Ma, as some contend (*9*), more than 35 percent of all the residues in these enzymes would have to be irreplaceable, a proposition we can reject on the basis of direct observation (*50*).

The residues in real proteins however, are not simply divided into those that change freely and those that do not change at all. Accordingly, we conducted an extensive simulation exercise to examine the impact of assigning every residue in a protein a specified probability for change (*51*). Not unexpectedly, the relationship between distance and similarity score becomes curvilinear under such circumstances. The impact on extrapolation is negligible, however, when distances are restricted to values corresponding to more than 30 percent sequence identity, only becoming significant when the similarity drops below 25 percent identity. Even when we assumed an extreme distribution of probabilities, the correction factor for a linear extrapolation to the eukaryote-eubacteria divergence time amounted to only 10 to 15 percent. In the end, a simple linear extrapolation has yielded a set of reasonable divergence times, especially when viewed in the light of offsetting if modest revisions required for observed differences in rate along different lineages.

In summary, our data show that, at least for the set of enzymes studied, eukaryotes and eubacteria last shared a common ancestor about 2 billion years ago, or twice as long ago as the existence of the last common ancestor of plants and animals (*52*). The estimate has survived critical assessment with regard to choice of weighting scale, random and selected data omission, changes in amino acid replacement rate along different lineages, and considerations having to do with the linear extrapolation of calculated distances. The magnitude and offsetting nature of these corrections suggest that the estimate is accurate to about 10 percent.

Amendments and extrapolations aside, the data indicate that bacterial sequences are more similar to each other than they are to their eukaryote counterparts. At first glance, this might seem to argue for a very early divergence of eukaryotes and eubacteria. But the common ancestor of prokaryotes and eukaryotes was already a very complex organism with a sophisticated and highly regulated metabolism; its genetic replicative machinery was very advanced and included most extant error-prevention devices. Moreover, during our casual inspection of enzyme candidates for this study, it was obvious that most bacterial

**Table 5.** Uncorrected eukaryote-eubacteria divergence times for sampled data sets*

| | PAM | BLOSUM |
|---|---|---|
| All 57 enzyme sets | 1.94 | 1.83 |
| Remove seven enzyme sets at random† | 1.92 | 1.81 |
| Remove seven slowest changers‡ | 1.98 | 1.86 |
| Remove seven fastest changers‡ | 1.83 | 1.69 |
| Remove seven lowest B/A ratios§ | 1.78 | 1.69 |
| Remove seven highest B/A ratios‖ | 2.10 | 1.97 |
| Use 27 randomly drawn | 2.12 | 1.96 |
| Use 27 remaining | 1.80 | 1.72 |
| Use 27 slowest changers‡ | 1.79 | 1.66 |
| Use 27 fastest changers‡ | 2.09 | 1.98 |

*These "divergence times" are actually the ratios of the eukaryote-eubacteria distance values (denoted B) divided by the animal-fungi distances (denoted A), but they are coincidentally about the same as the time in billions of years. †Average of 10 trials. ‡As determined by the animal-fungi distance (A). §Those entries with the lowest B/A ratios would be the ones most likely to be horizontal imports. ‖Those entries with the highest B/A ratios would be the ones most likely to be paralogs.

enzymes have orthologous or paralogous homologs among the eukaryotes. If living organisms existed as much as 3500 Ma and the last common ancestor of prokaryotes and eukaryotes lived about 2000 Ma, then there would have been 1500 million years for this finely tuned and complex arrangement to evolve.

However, if all extant bacteria date back to a common ancestor less than 2 billion years ago, questions must be asked as to what kind of organism gave rise to the present bacterial kingdom and what kinds of creatures existed before that time. Whether all but one of the early lineages of bacteria became extinct and other similar questions require addressing in the light of the determined chronology (*53*).

## REFERENCES AND NOTES


1. J. W. Schopf, *Science* **260**, 640 (1993).
2. A. H. Knoll, *ibid*. **256**, 622 (1992).
3. E. Zuckerkandl and L. Pauling, in *Evolving Genes and Proteins*, V. Bryson and H. J. Vogel, Eds. (Academic Press, New York, 1965), pp. 97–166.
4. R. F. Doolittle and B. Blomback, *Nature* **202**, 147 (1964).
5. As an example, an entire issue of *J. Mol. Evol.* [**26**, 1–164 (1987)] was devoted to molecular clocks and their use in different settings.
6. T. H. Jukes, *Space Life Sci.* **1**, 469 (1969).
7. P. J. McLaughlin and M. O. Dayhoff, *Science* **168**, 1469 (1970); M. Kimura and T. Ohta, *Nature New Biol*. **243**, 199 (1973); H. Hori and S. Osawa, *Proc. Natl. Acad. Sci. U.S.A.* **76**, 381 (1979); M. O. Dayhoff, *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Washington, DC, 1978), vol. 5, suppl. 3.
8. J. W. Schopf and D. Z. Oehler, *Science* **193**, 47 (1976).
9. C. R. Woese, in *Evolution from Molecules to Men*, D. S. Bendall, Ed. (Cambridge Univ. Press, London, 1983), pp. 209–233; M. L. Sogin, in *New Perspectives on Evolution*, L. Warren and H. Koprowski, Eds. (Wiley-Liss, New York, 1991), pp. 175–188.
10. Long-term amino acid replacement in proteins is subject to a large number of independently acting factors, each of which may have only a small effect on the overall process. In our study we have assumed that the replacement process is approximately uniform, an assumption that is open to question [see J. H. Gillespie, *The Causes of Molecular Evolution* (Oxford Univ. Press, New York, 1991)].
11. R. F. Doolittle, K. L. Anderson, D. F. Feng, in *The Hierarchy of Life*, B. Fernholm, K. Bremer, H. Jornvall, Eds. (Elsevier, Amsterdam, 1989), pp. 73–85.
12. Some enzymes that have the same name and Enzyme Commission (E.C.) numbers, but are derived from independent origins, include superoxide dismutases, carbonic anhydrases, aldolases, and serine proteases.
13. W. M. Fitch, *Syst. Zool.* **19**, 99 (1970).
14. M. W. Smith, D. F. Feng, R. F. Doolittle, *Trends Biochem. Sci.* **17**, 489 (1992).
15. Most of the 531 amino acid sequences used in our study were taken from Release 42 (30 September 1994) of the Protein Identification Resource (PIR) although some sequences identified recently were abstracted from GenBank. First, a list was compiled of all entries in the PIR that included official E.C. numbers (*16*). Then a tally was made of how many different entries were listed for each E.C. number. Any enzyme with four or more entries was examined to see whether at least three major groups were represented (animal, plants or fungi, and eubacteria); if so, the enzyme was considered a possible candidate for inclusion in the study. All told, Release 42 of the PIR contained 13,653 entries with E.C. identification numbers. Of these, 1262 E.C. numbers were present, accounting for just under 40 percent of the officially declared 3196 enzymes (*16*). About half of these had three entries or fewer and were not considered further. The half with four or more entries was screened with regard to organismic representation. Sequences for enzymes encoded by organellar DNA (mitochondria and chloroplasts) and sequences from viruses were not included. The sequences of candidate groups were aligned and phylogenies were constructed (*17–22*). If the phylogenetic trees seemed reasonable, by which we mean there was no evidence of horizontal gene transfer or adulteration by paralogous comparisons (*23*), the sequence subset became a part of the study. The entire set (divided into the six standard enzyme groups) can be obtained by anonymous ftp from juno.ucsd.edu. cd to directory pickup.
16. *Enzyme Nomenclature, Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology* (Academic Press, New York, 1992).
17. Candidate groups were aligned by the progressive method (*18*), a procedure that uses the Needleman-Wunsch algorithm (*19*). Several different substitution matrices were used, including the Dayhoff PAM-250 (*20*), GCB matrix (*21*), and the BLOSUM-62 matrix (*22*). The GCB comparisons were not significantly different from those obtained with the Dayhoff PAM-250 scale, and those results have not been included in this study. The BLOSUM-62 scale, however, resulted in obviously improved alignments for the most distant of the relationships. We therefore used it to obtain all the final alignments, even though we then used the PAM-250 table to calculate distances for comparison with those obtained from the BLOSUM table. The comparison data were conveniently managed with the aid of the commercially available Microsoft Excel spreadsheet software. Entry sheets listing species represented, lengths of sequences, and such items were prepared for each of the 57 enzymes, as were other sets of primary data sheets that included all resemblances and distances between groups. Summary "charts" of distances and percent identities were prepared from the entire data set or from designated subsets (Table 4). As new data become available, it is relatively easy to update the records and recalculate all values.
18. D. F. Feng and R. F. Doolittle, *J. Mol. Evol.* **25**, 351 (1987); *Methods Enzymol.* **183**, 375 (1990).
19. S. B. Needleman and C. D. Wunsch, *J. Mol. Biol.* **48**, 443 (1970).
20. R. M. Schwartz and M. O. Dayhoff in *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, Washington, DC, 1978), vol. 5, supp. 3, pp. 353–358.
21. G. H. Gonnet, M. A. Cohen, S. A. Benner, *Science* **56**, 1443 (1992).
22. S. Henikoff and J. G. Henikoff, *Proteins: Struct. Funct. Genet.* **17**, 49 (1993).
23. Several situations in which displacement during evolution has led to functional convergence after paralogous radiations include bacterial ornithine decarboxylase (E.C. 4.1.1.17), which is obviously more like lysine decarboxylase (E.C. 4.1.1.18) than it is to eukaryotic ornithine decarboxylases, and bacterial tyrosine transaminase (E.C. 2.6.1.5), which is more similar to bacterial aspartate transminase (E.C. 2.6.1.1) than it is to the eukaryotic tyrosine enzyme. Other enzyme sets that were not included on the basis of anomalous phylogenetic trees were catalase (E.C. 1.11.16), pyrroline carboxylate reductase (E.C. 1.5.1.2), glutathione reductase (E.C. 1.6.4.2), phosphoribosylglycineamide formyl transferase (E.C. 2.1.2.2), transketolase (E.C. 2.2.1.1), glycogen phosphorylase (E.C. 2.4.1.1), hypoxanthine transferase (E.C. 2.4.2.8), orotate phosphate transferase (E.C. 2.4.2.10), glutathione transferase (E.C. 2.5.1.18), galactokinase (E.C. 2.7.1.6), adenylate kinase (E.C. 2.7.4.3), uridine 5′-diphosphate–glucose-hexose phosphate uridyl transferase (E.C. 2.7.7.9 and E.C. 2.7.7.12), ornithine decarboxylase (E.C. 4.1.1.17), glucose phosphate isomerase (E.C. 5.3.1.9), and phosphoglycerate mutase (E.C. 5.4.2.1).
24. R. F. Doolittle, in *Methods in Protein Sequence Analysis*, K. Imahori and F. Sakiyama, Eds. (Plenum, New York, 1993), pp. 241–246.
25. Average percent identities notwithstanding, the data were not entirely consistent. Of the nine possible comparisons, in five cases the archaebacterial sequences clustered with the eukaryotes, and in three with the eubacteria. In one case (phosphoglycerate kinase, E.C. 2.7.2.3) the eubacteria and eukaryote sequences were more similar to each other than to the archaebacterial sequence.
26. D. F. Feng, M. S. Johnson, R. F. Doolittle, *J. Mol. Evol.* **21**, 112 (1985).
27. The use of the Poisson distribution as a probabilistic model for amino acid replacement dates back to Zuckerkandl and Pauling (*3*). It is often used in the simple form $D = -\ln(1 - p/n)$, with $p/n$ being the fraction of changed residues. In this form, the equation mainly corrects for the unobserved occurrence of two or more replacements at the same site (*28*). Numerous modifications have been reported, including attempts to correct for invariant residues (*29, 30*) or chance occurrences (*31, 26*).
28. For example, M. Kimura and T. Ohta, *J. Mol. Evol.* **1**, 1 (1971); R. E. Dickerson, *ibid*., p. 26; M. Kimura and T. Ohta, *ibid*. **2**, 87 (1972).
29. E. Margoliash and E. Smith, in *Evolving Genes and Proteins*, V. Bryson and H. J. Vogel, Eds. (Academic Press, New York, 1965), pp. 221–242.
30. W. M. Fitch and E. Markowitz, *Biochem. Genet.* **4**, 579 (1970).
31. T. H. Jukes and C. R. Cantor, in *Mammalian Protein Metabolism*, H. N. Munro, Ed. (Academic Press, New York, 1969), pp. 21–132.
32. R. Johnson, *Elementary Statistics* (Duxbury Press, Boston, 1984).
33. C. R. Marshall, *J. Mol. Evol.* **30**, 400 (1990); M. J. Benton, *ibid*., p. 409; R. L. Carroll, *Vertebrate Paleontology and Evolution* (Freeman, New York, 1988).
34. S. C. Conway Morris, *Nature* **361**, 219 (1993).
35. The "relative rate test" [V. M. Sarich and A. C. Wilson, *Science* **179**, 1144 (1973)] can be used whenever the intergroup distances for three taxa (or more) are available. For example, it was clear from a consideration of the fungi-eubacteria and animal-eubacteria distances that fungal sequences were changing faster than animal ones. We were able to apply this simple test to all the taxa in our study.
36. S. L. Baldourf and J. D. Palmer, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 11558 (1993). P. O. Wainright, G. Hinkle, M. L. Sogin, S. K. Stickel, *Science* **260**, 340 (1993).
37. A. H. Knoll, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 6743 (1994).
38. M. L. Sogin, J. H. Gunderson, H. J. Elwood, R. A. Alonso, D. Q. A Peattie, *Science* **243**, 75 (1989).
39. T. Cavalier-Smith, *Microbiol. Rev.* **57**, 953 (1993).
40. W. F. Loomis and D. W. Smith, *Proc. Natl. Acad. Sci. U.S.A.* **87**, 9093 (1990).
41. Z. Zhang, *J. Micropalaeontol*, **5**, 9 (1986).
42. T.-H. Han and B. Runnegar, *Science* **257**, 232 (1992).
43. N. Iwabe, K.-I. Kuma, M. Hasegawa, S. Osawa, T. Miyata, *Proc. Natl. Acad. Sci., U.S.A.* **86**, 9355 (1989).
44. H. Ochman and A. C. Wilson, *J. Mol. Evol.* **26**, 74 (1987).
45. Matrices of pairwise distances were examined by the program BLEN (*18*) which uses a least squares approach (*46*).
46. L. C. Klotz and R. L. Blanken, *J. Theoret. Biol.* **91**, 261 (1981).
47. C. H. Langley and W. M. Fitch, *J. Mol. Evol.* **3**, 161 (1974). W. M. Fitch and C. H. Langley, *Fed. Proc.* **35**, 2092 (1976). W. M. Fitch and F. J. Ayala, *Proc. Natl. Acad. Sci. U.S.A.* **91**, 6802 (1994).
48. An extreme but valid example is the case of enzyme sequences in retroviruses. For example, a survey of 26 ribonuclease H sequences revealed that only 4 of 120 residues remained unchanged; R. F. Doolittle, D-F. Feng, M. S. Johnson, M. A. McClure, *Q. Rev. Biol.* **64**, 1 (1989).
49. M. Nei, *Molecular Evolutionary Genetics* (Columbia Univ. Press, New York, 1987).
50. Although, on the average, eukaryote and eubacteria sequences are 37 percent identical, the observed fraction of irreplaceable residues was, again on average, only 17 percent. There was also a natural tendency for the fraction of irreplaceable

residues to be smaller; the larger the number of sequences in a set, the extrapolated fraction being about 5 percent.
51. A computer model has been constructed that follows the divergence of mutated protein sequences under various circumstances of constraint (R. F. Doolittle and D. F. Feng, in preparation).
52. One of the earliest estimates made about the prokaryote-eukaryote divergence concluded, on the basis of a relatively small number of transfer RNA sequences, that the split occurred about twice as long ago as the divergence of plants, animals, and fungi (*6*).
53. There will be some who will remind us of alternative scenarios concerning the origin of eukaryotic organisms, and especially of the possibility that some of the sequences discussed here were actually imported by an archaebacterial symbiont destined to become the nucleus. The fusion of a eubacterial "prokaryote" and an archaebacterium has been widely discussed (*54*). Although we are skeptical of such models on other grounds, we should point out that such an occurrence would not affect our findings, except that the time we are reporting as a divergence time for eukaryotes and eubacteria would instead chronicle the alleged fusion event.
54. R. S. Gupta and G. B. Golding, *J. Mol. Evol.* **37**, 573 (1993); G. B. Golding and R. S. Gupta, *Mol. Biol. Evol.* **12**, 1 (1995).
55. We thank K. Anderson for assistance in preparing this manuscript and S. Frank, J. Gillespie, and two anonymous reviewers for helpful suggestions. Supported in part by NIH grant HL-26873.

RESEARCH ARTICLE

# Thiyl Radicals in Ribonucleotide Reductases

Stuart Licht, Gary J. Gerfen, JoAnne Stubbe

The ribonucleoside triphosphate reductase (RTPR) from *Lactobacillus leichmannii* catalyzes adenosylcobalamin (AdoCbl)-dependent nucleotide reduction, as well as exchange of the 5′ hydrogens of AdoCbl with solvent. A protein-based thiyl radical is proposed as an intermediate in both of these processes. In the presence of RTPR containing specifically deuterated cysteine residues, the electron paramagnetic resonance (EPR) spectrum of an intermediate in the exchange reaction and the reduction reaction, trapped by rapid freeze quench techniques, exhibits narrowed hyperfine features relative to the corresponding unlabeled RTPR. The spectrum was interpreted to represent a thiyl radical coupled to cob(II)alamin. Another proposed intermediate, 5′-deoxyadenosine, was detected by rapid acid quench techniques. Similarities in mechanism between RTPR and the *Escherichia coli* ribonucleotide reductase suggest that both enzymes require a thiyl radical for catalysis.

Although the reactivity of free radicals has often been associated with mutagenesis and molecular degradation, sophisticated methods have evolved to harness this reactivity to effect difficult reactions with remarkable selectivity. The past few years have witnessed a renaissance in the detection of protein-derived radicals that have been proposed to play essential roles in metabolism, from DNA biosynthesis and repair to prostaglandin biosynthesis and acetyl–coenzyme A production (*1*–*4*). The *Escherichia coli* ribonucleotide reductase (RNR), which has served as a prototype for these enzymes, was demonstrated, in 1977, to contain a stable tyrosyl radical that plays an essential role in the conversion of all nucleotides to deoxynucleotides (*5*). This reduction is accompanied by oxidation of two cysteines to a disulfide (Scheme 1), and additional turnovers require re-reduction of the enzyme by a reducing system such as thioredoxin (TR), thioredoxin reductase (TRR), and nicotinamide adenine dinucleotide phosphate reduced (NADPH) (Scheme 1) (*6*). Ribonucleotide reductases, despite their central role in deoxynucleotide formation in all organisms, have been shown over the past decades to contain metallo-cofactors that are structurally and chemically distinct (Fig. 1) (*7*–*9*). The reductase from *Lactobacillus leichmannii* requires adenosylcobalamin (AdoCbl) as a cofactor, which can generate cob(II)alamin and a putative 5′-deoxyadenosyl radical (5′-dA·) in a kinetically competent fashion (*10*, *11*). The reductase from *E. coli* grown under anaerobic conditions uses an iron-sulfur cluster and *S*-adenosylmethionine to generate a glycyl radical essential for nucleotide reduction (*12*), and a reductase from *Brevibacterium ammoniagenes* uses a manganese cluster to generate a putative protein radical (*13*). All of these reductases are associated with metallo-cofactors that are thought to generate, in the protein environment, an organic radical that initiates the nucleotide reduction process. However, in no case has a protein radical in a reductase system been demonstrated to disappear and reappear with a rate faster than the turnover of the enzyme (*7*).

The two reductases whose mechanisms have been examined in the greatest detail are those from *E. coli* and *L. leichmannii*. Even though each of these proteins possesses a characteristic primary and quaternary structure and a distinct metallo-cofactor, an in-depth examination of these proteins with mechanism-based inhibitors and site-directed mutants has revealed an extensive congruence in catalytic details (*7*, *8*, *14*). The role of the metallo-cofactor appeared to be even more complex than originally hypothesized, and, in 1990, the proposal was made that the function of the tyrosyl radical in the *E. coli* reductase and the putative 5′-dA· in the *L. leichmannii* reductase was to generate a thiyl radical, which initiated the nucleotide reduction process by abstraction of the 3′ hydrogen atom from the nucleotide substrate (*7*, *8*). Direct evidence in support of this proposal, however, has remained elusive.

We now describe the direct evidence for the intermediacy of a thiyl radical ($C^{408}$) in the nucleotide reduction process catalyzed by the *L. leichmannii* reductase. Even though there is no statistically significant sequence similarity between the *E. coli* and the *L. leichmannii* reductases (*15*), the sequence context surrounding the putative

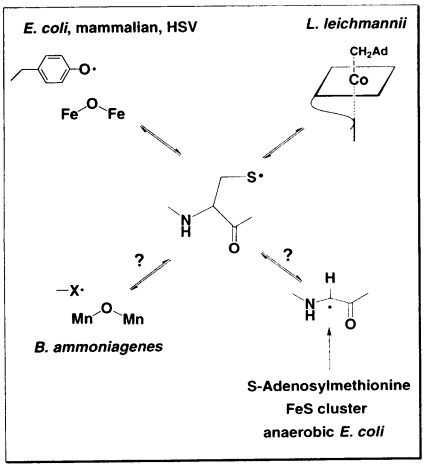


**Fig. 1.** Metallo-cofactors of RNRs required for the generation of the putative thiyl radical essential for the nucleotide reduction process.

S. Licht is in the Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. G. J. Gerfen is at the Francis Bitter Magnet Laboratory and is in the Department of Chemistry, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. J. Stubbe is in the Departments of Chemistry and Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.