Determining Divergence Times

different proteins comprising 531 different

On the average, plant sequences are more like animal sequences than are fungal ones (Fig. 1, A and B). This is true whether

ensure against some hidden bias. In all cases the similarity scores obtained were scaled as follows (26):

aligned amino acids obtained from the weight matrices, S_{rand} the corresponding score for two random sequences of the same 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.

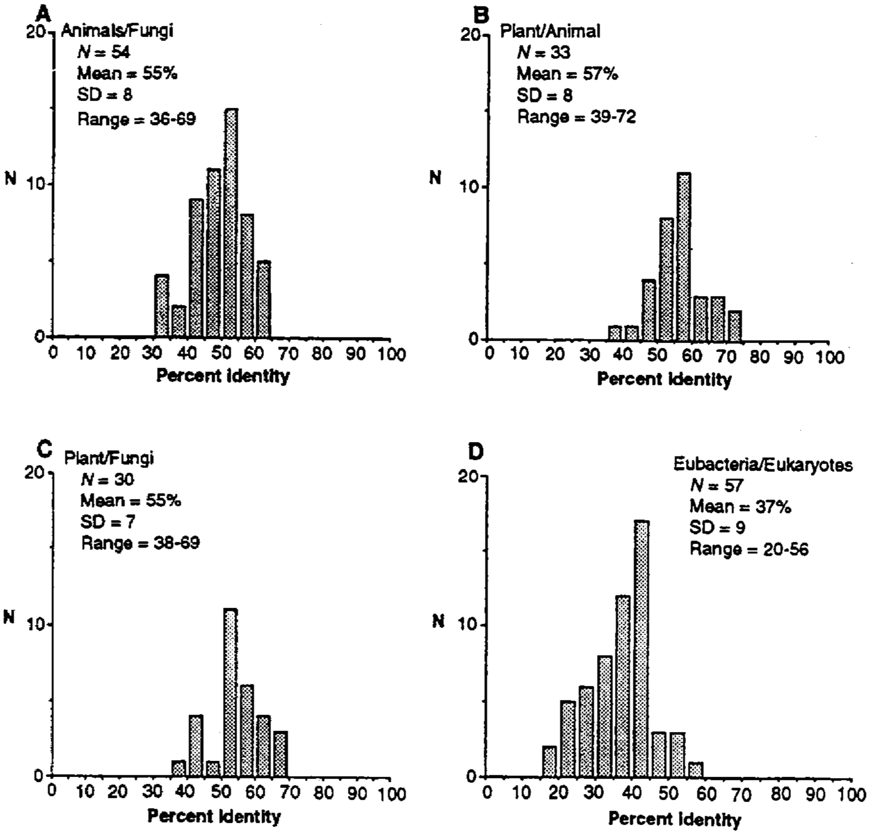

Fig. 1. Resemblances (percent identity) of enzyme sequences from principal biological groups as measured in blocks of five percentage points.

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.

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

indicates that eukaryotes last shared a common ancestor with archaebacteria 1800 Ma,

to 1900 million years and that resemble eukaryotic cells (41), but they are at odds with

[redacted]

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 anoma-

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	12	1.06

not every taxon was represented. Again, our

values are the more rigorously determined

and eubacteria by only 10 percent, barely 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

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

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