MOLECULAR PHYLOGENETICS AND EVOLUTION
Vol. 6, No. 2, October, pp. 189–213, 1996
ARTICLE NO. 0071

Reconstruction of Ancient Molecular Phylogeny

RODERIC GUIGÓ,*,†,‡ ILYA MUCHNIK,‡ AND TEMPLE F. SMITH‡,[1]

Theoretical Biology and Biophysics, T-10, MS K710, LANL, Los Alamos, New Mexico 87545; †Informàtica Mèdica, Institut Municipal d'Investigació Mèdica, C/Dr. Aiguader 80, 08003 Barcelona, Spain; and ‡BioMolecular Engineering Research Center, College of Engineering, Boston University, 36 Cummington Street, Boston, Massachusetts 02215

Received March 25, 1996

Support for contradictory phylogenies is often obtained when molecular sequence data from different genes is used to reconstruct phylogenetic histories. Contradictory phylogenies can result from many data anomalies including unrecognized paralogy. Paralogy, defined as the reconstruction of a phylogenetic tree from a mixture of genes generated by duplications, has generally not been formally included in phylogenetic reconstructions. Here we undertake the task of reconstructing a single most likely evolutionary relationship among a range of taxa from a large set of apparently inconsistent gene trees. Under the assumption that differences among gene trees can be explained by gene duplications, and consequent losses, we have developed a method to obtain the global phylogeny minimizing the total number of postulated duplications and losses and to trace back such individual gene duplications to global genome duplications. We have used this method to infer the most likely phylogenetic relationship among 16 major higher eukaryotic taxa from the sequences of 53 different genes. Only five independent genome duplication events need to be postulated in order to explain the inconsistencies among these trees.

1. INTRODUCTION

Phylogenetic trees inferred from the sequence of different genes are often contradictory (see for example, Field *et al.*, 1988; Lake, 1990). Inconsistencies in extrapolating species trees from molecular sequence data often arise from unrecognized paralogy, that is, duplication of genes before separation of lineages (Forterre *et al.*, 1993; Hasegawa and Hashimoto, 1993; Sogin *et al.*, 1993). Gene duplication may have occurred frequently during eukaryotic history and may have been a common mechanism through which new genes originated (Ohno, 1970). Indeed the entire genome of the higher eukaryotic taxa has apparently undergone a number of near complete duplications (Rees and Jones, 1972; Grime and Mowforth, 1982). Genome duplications—total or partial—would create a number of potentially independently evolving lines of descent in which duplicated genes could have been differentially lost or evolved to different functions. Phylogenetic trees inferred from genes not recognized as ancient duplicates may fail to reflect the true phylogenetic relationships among the species.

In this paper we present a formal method for constructing a global species phylogeny from a set of (potentially inconsistent and partial) gene trees. Potential inconsistencies among trees are assumed for the purpose of this study to be only the consequence of gene duplications. The global phylogeny is built on the parsimonious criterion of minimizing the number of gene duplications that need to be postulated to reconcile the individual gene trees into the global phylogeny. The method is based on the concept of mapping between a gene tree and species tree, first introduced by Goodman *et al.* (1979) and recently discussed by Page (1994) within the framework of "reconciled trees." Within a formalization of Goodman's mapping, we derive definitions for the concept of gene duplication, which is essentially equivalent to the procedural definition given in Goodman *et al.* (1979). The resulting formalism allows us to compute the number of gene duplications and losses needed to map a gene tree into a species tree. This number can be assumed to be a measure of dissimilarity between gene and species tree, and it is related to the dissimilarity measure used in Goodman *et al.* However while Goodman *et al.* use such a measure as a minimization criterion for mapping potential gene trees into an a priori assumed species tree, we use it here as the minimization criterion for finding the species tree into which a number of known gene trees have to be mapped. Thus, while the goal of Goodman *et al.* is to find the "best" one gene tree for a known species phylogeny, our goal is to find the "best" species phylogeny for a set of known gene trees. To obtain such a phylogeny, we use a nearest neighbor branch swapping algorithm (Waterman and Smith, 1978). Given a set of (possibly inconsistent and partial) gene trees, the algo-

[1] To whom correspondence should be addressed. Fax: (617) 353-7020. E-mail: tsmith@darwin.bu.edu.

1055-7903/96 $18.00

rithm finds the global species phylogeny minimizing the total number of gene duplications and losses needed to map the gene trees into such a phylogeny.

Second, we present a method to map gene duplications from the individual gene trees into the obtained species tree and to cluster such mapped duplications into a few "genome duplications." Again the criterion to derive such a map is a parsimonious one: minimizing the number of locations in the species tree where the gene duplications are mapped. To cluster gene duplications in the species tree into genome duplications, we assume that gene duplications occurring at the same location in the species tree are the result of the same genome duplication.

To illustrate the above methodology we have attempted to reconstruct the phylogenetic relationships among 16 major higher eukaryotic taxa from a large set of individual gene trees. Although resolving the actual eukaryotic phylogeny is not our primary goal, we have taken pains to obtain a robust and reliable set of gene trees (i.e., a set of trees whose topology is very likely to properly reflect the sequence similarity relationships). Thus the inconsistencies among the trees have a reasonable probability of resulting from paralogy rather than from mistakes in the process of inferring the tree topology from the sequence data. This effort has led to a generally useful procedure, which is essentially based on aligning only the regions of unquestionable homology to compute the distances between the sequences, and considering only those cases for which maximal and minimal linkage result in the same topology when inferring the tree from the sequence distances. After such a rigorous filtering, 53 different genes trees were obtained and used in our analysis. They were highly contradictory—to an unexpected degree given the extent to which molecular sequence data are being used to infer phylogenetic relationships between species. A global phylogeny was obtained from the 53 genes trees using the methodology developed here. The phylogeny is reasonable from a biological standpoint and, interestingly, requires only five major genome duplications during the eukaryotic history to explain the observed contradictions among the individual gene trees.

In what follows, we first describe the generation of a large set of individual gene trees, since it was the observation that they are highly contradictory that motivated us to investigate the problem addressed in this paper. Second, we describe the formalization of the concept of map between gene and species tree and define within this formalization the concepts of gene duplication and loss. Next, we describe an algorithm to find the species tree globally that is consistent with a set of individual gene trees. This is then applied to the set of actual gene trees. Next, we describe a method to map and cluster individual gene duplications into probable genome duplications in the global species tree. Finally, as a proof of principle, we apply the latter method to postulate and place a minimum number of genome duplications in the phylogeny obtained for the major eukaryotic divisions.

TABLE 1

The Taxonomic Groups Considered Following the Nomenclature Used in the SWISSPROT Databases

Taxon	Number of genes	Number of sequences
Acoelomates	16	17
Agnatha	6	8
Amphibia	55	74
Annelida	7	11
Arthropoda	56	112
Aves	51	130
Chondrichthyes	18	27
Chlorophyceae	23	36
Echinodermata	35	62
Embryophyta	57	173
Fungi	61	148
Mammalia	93	589
Mollusca	12	20
Osteichthyes	28	75
Protozoa	45	94
Reptilia	11	34

Note. 106 gene families were obtained for which sequences from species belonging to at least four of these groups existed in SWISSPROT, release 19. For each taxonomic group, the second column is the number of families for which there is a sequence for at least one species in the group. The third column is the total number of sequences for the taxonomic group belonging to the 106 gene families.

2. THE SEQUENCE DATA

Sixteen major higher eukaryotic taxonomic groups were considered (see Table 1). These taxa were chosen because, with a few exceptions, they are recognized as natural clades—that is, the species in these taxa are believed to share a common ancestor that is not shared by any other species outside the taxa. Such an assumption is probably wrong for protozoa and reptilia and has been questioned for other taxa, such as agnatha (for example, Hardisty, 1979) and arthropoda (Lake, 1990). The protein sequences obtained from species belonging to these groups were extracted from SWISSPROT, release 19. Assignment of species to taxa was done following SW taxonomic classification; 12,272 sequences were obtained corresponding to 5892 different genes. Initial assignment of sequences to genes was done following SWISSPROT nomenclature and annotation. In the SWISSPROT database, the prefix of the locus name for a given sequence denotes the gene to which the protein sequence corresponds. Sequences sharing the same prefix in their locus names were initially considered as corresponding to the same gene. Gene assignments were confirmed by the inspection of the definition field

in the corresponding SWISSPROT entry. Thus our initial gene definitions were based on a common function and/or recognized sequence similarity as recorded in the swissprot database. All partial and/or fragment gene sequences were removed. Only those genes for which sequences were available from at least four different taxonomic groups were further considered. This resulted in a set of 1597 protein sequences corresponding to 104 different genes. All 16 eukaryotic taxa are represented in this set, which contains sequences from 501 different eukaryotic species. The number of genes and the number of sequences obtained for each taxa appear in Table 1. The genes, as well as the number of sequences and the number of taxa for each of the genes, appear in Table 2.

3. TREES OBTAINED FROM DIFFERENT GENE SEQUENCES ARE CONTRADICTORY

For each one of the above 104 genes, the corresponding sequences from the different species were aligned using a pattern-induced multiple alignment algorithm (Smith and Smith, 1992). The algorithm is an optimal local dynamic program (Smith and Waterman, 1981) that aligns a set of homologous sequences based on their common pattern of conserved sequence elements. In order to identify statistically unquestionable aligned regions, the Information Content of each pattern (Smith and Smith, 1992) was computed. Only genes were kept for which the common pattern relative Information Density was greater than 0.3 (equivalent to 30% amino acid identities) or whose pattern relative Information Density was at least 0.25 and the equivalent of 45 amino acid identities. Sixty-seven gene families passed this test. The resulting alignment for each of these genes was inspected by hand and, in some cases modified by minor alternate gap placement to increase local Information Density. (The common pattern information density was not the aligning optimization criterion.)

Pairwise distances between sequences (species) for each gene were derived from the above multiple local alignments. The pairwise distances between the sequences (species) for each gene family were computed only over regions common to all the sequences. In addition, all positions in these regions were discarded for which at least one of the sequences required an alignment gap. A simple distance was computed: the total number of estimated nucleotide differences between the amino acid sequences. This estimated number of nucleotide differences between amino acid positions was the minimum number of nucleotides required to map one amino acid into another.

For each gene, pairwise distances between taxa were derived from the above sequence distances. The distance between two taxa is computed as the average of all pairwise distances between sequences belonging to one taxon and sequences belonging to the other. That is, if A and B are taxa with N_A and N_B sequences, respectively, the distance between A and B, d_{AB} is the average distance

$$d_{AB} = \frac{1}{N_A N_B} \sum_{i \in A} \sum_{j \in B} d_{ij},$$

where d_{ij} is the distance between sequences (species) i from taxa A and j from taxa B.

Thus for each gene, a 16×16 distance matrix was obtained, with at least six nonempty elements (corresponding to the pairwise distances among at least four taxa). From these matrices, rooted trees were initially obtained using the Fitch and Margoliash procedure (Fitch and Margoliash, 1967), with the root being placed assuming that the total length from the root to any terminal leaf is nearly the same. To identify topologically ill-defined trees, two additional distance clustering procedures were employed: maximal linkage and single linkage. Only those genes for which maximal and single linkage resulted in the same clustering were further considered. Note that if both maximum and single linkage generate identical branching topologies, it is unlikely that other distance-based binary clustering will give a different topology. The Fitch and Margoliash trees for the remaining 53 genes satisfying this condition appear in Fig. 1. These trees are our final set of gene trees.

As shown in Fig. 1, even trees obtained under these relatively restrictive conditions can be highly contradictory (even considering only tree topology or branching order and not branch lengths). Thus for example, while the hypothesis that echinodermata and vertebrates (chordata) are members of a monophyletic group is supported by the gene trees obtained from the sequences of some histones (H1B, H3), enzymes (HMDH), and tubulins (TBB), the trees obtained from the sequences of other histones (H2B) and some actins (ACT2 and ACT3) suggest a closer phylogenetic relationship between echinodermata and the other invertebrates. According to the gene tree derived from the tubulin α-1 sequence, arthropoda are phylogenetically closer to vertebrates than echinodermata. Similarly, the individual gene trees obtained are contradictory on the issue of the recently suggested evolutionary link between fungi and metazoa (Wainright *et al.*, 1993). According to the majority of trees, metazoa and fungi share a more recent common ancestor than either does with the protozoa. The trees derived from some tubulins (TBA1, TBB, TBB1), ubiquitin (UBIQ), and a ribosomal protein (RLA1), on the other hand, suggest that fungi diverged before the splitting of protozoa and metazoa. In this case, however, inconsistency may be compounded by the (wrong) assumption that the proto-

TABLE 2

The 104 Gene Families Initially Considered

		Gene	Number of taxa	Number of sequences	Length of alignment	IC	ID
AACT	*	α-actinin	4	4	865	455.55	0.53
ACH2	*	Acetylcholine receptor protein α-2 chain	4	4	456	220.77	0.48
ACHD	*	Acetylcholine receptor protein δ chain	4	5	512	320.55	0.63
ACHG	*	Acetylcholine receptor protein γ chain	4	6	507	308.84	0.61
ACT	*	Actin	5	17	354	207.23	0.59
ACT1		Actin 1	6	13	375	259.48	0.69
ACT2	*	Actin 2	6	10	376	262.41	0.70
ACT3	*	Actin 3	5	6	373	310.03	0.83
ACTB	*	Actin B	4	4	374	370.07	0.99
ADH		Alcohol dehydrogenase	4	14	236	31.07	0.13
ADH1		Alcohol dehydrogenase I	4	14	245	28.96	0.12
ALF		Fructose-biphosphate aldolase	4	7	337	56.20	0.17
ANF	*	Atrial natriuretic factor	4	10	23	14.80	0.64
ANFC	*	C-Type natriuretic peptide	4	4	22	18.11	0.82
ATP6		ATP Synthase A chain	7	21	226	35.59	0.16
ATP8		ATP Synthase protein 8	8	20	13	2.61	0.20
ATPB	*	ATP Synthase β chain	4	15	466	323.86	0.69
CALM		Calmodulin	9	22	146	89.52	0.61
CAP1		Calpain I	4	4	307	37.81	0.12
CATA	*	Catalase	4	8	467	122.03	0.26
CISY	*	Citrate synthase	4	4	431	239.70	0.56
COLI	*	Corticotropin	5	14	39	24.24	0.62
COX1		Cytochrome c oxidase polypeptide I	9	27	482	150.33	0.31
COX2		Cytochrome c oxidase polypeptide II	8	30	111	13.01	0.12
COX3		Cytochrome c oxidase polypeptide III	9	25	255	43.52	0.17
CRAA	*	α-Crystallin A chain	4	36	169	121.47	0.72
CYB		Cytochrome b	9	23	338	46.32	0.14
CYC		Cytochrome c	16	80	97	30.48	0.31
CYLA	*	Cyclin A	4	4	394	155.24	0.39
CYLB	*	Cyclin B	4	6	359	118.60	0.33
CYPH	*	Peptidyl-prolyl *cis-trans* isomerase	4	12	167	74.31	0.44
EF1A		Elongation factor 1-α	6	11	442	311.46	0.70
G3P	*	Glyceraldehyde 3-phosphate dehydrogenase	5	12	328	196.72	0.60
G3P1	*	Glyceraldehyde 3-phosphate dehydrogenase 1	4	4	332	218.38	0.66
G3P2	*	Glyceraldehyde 3-phosphate dehydrogenase 2	4	4	331	226.79	0.69
G6PI	*	Glucose-6-phosphate isomerase	4	8	537	217.98	0.41
GAP1	*	GAP Junction α-1 protein	4	6	374	115.07	0.31
GLB		Globin (myoglobin)	4	10	104	5.77	0.06
GLB1		Globin I	6	13	95	0.85	0.01
GLB2		Globin IIB	4	6	129	13.89	0.11
GLUC	*	Glucagon	4	13	30	21.30	0.71
H1		Histone H1	6	12	143	15.38	0.11
H1B	*	Histone H1B	4	4	121	39.63	0.33
H2A		Histone H2A	10	15	120	91.86	0.77
H2A1		Histone H2A.1	6	7	122	84.70	0.69
H2A2	*	Histone H2A.2	6	7	123	85.62	0.70
H2A3	*	Histone H2A.3	4	4	117	92.94	0.79
H2B	*	Histone H2B	8	12	122	96.56	0.79
H2B1	*	Histone H2B.1	5	9	118	74.66	0.63
H2B2		Histone H2B.2	6	9	118	71.84	0.61
H3	*	Histone H3	6	11	134	117.76	0.88
H31	*	Histone H3.1	4	4	134	117.69	0.88
H32	*	Histone H3.2	4	5	134	121.81	0.91
H4	*	Histone H4	8	11	103	85.11	0.83
HBA		Hemoglobin α-A chain	6	161	144	22.64	0.16
HBA1	*	Hemoglobin α-1 chain	5	13	142	45.61	0.32
HBA2	*	Hemoglobin α-2 chain	4	9	141	47.10	0.33
HBB		Hemoglobin β chain	6	151	133	27.13	0.20
HBB1		Hemoglobin β-1 chain	5	11	141	41.33	0.29
HBB2		Hemoglobin β-2 chain	4	11	140	40.91	0.29
HMDH	*	3-Hydroxy-3-methylglutaryl-coenzyme A reductase	6	8	525	192.00	0.37

TABLE 2—*Continued*

		Gene	Number of taxa	Number of sequences	Length of alignment	IC	ID
HS70		Heat shock 70 kDa protein	6	10	633	369.11	0.58
IGF1		Insulin-like growth factor I	4	7	70	59.03	0.84
INS		Insulin	6	43	47	21.02	0.45
LEG		D-Galactoside-specific lectin	4	5	98	19.07	0.19
MT1		Metallothionein-I	6	13	20	3.88	0.19
MT2		Metallothionein-II	5	14	32	4.03	0.13
MYG		Myoglobin	5	72	147	39.88	0.27
NGF	*	β-Nerve growth factor precursor	4	8	117	85.42	0.73
NU1M		NADH-Ubiquinone oxidoreductase chain 1	9	17	227	28.17	0.12
NU2M		NADH-Ubiquinone oxidoreductase chain 2	8	13	185	9.58	0.05
NU3M		NADH-Ubiquinone oxidoreductase chain 3	9	22	94	14.37	0.15
NU4M		NADH-Ubiquinone oxidoreductase chain 4	8	15	417	68.12	0.16
NU5M		NADH-Ubiquinone oxidoreductase chain 5	8	15	511	94.69	0.19
NU6M		NADH-Ubiquinone oxidoreductase chain 6	7	12	147	16.12	0.11
NULM		NADH-Ubiquinone oxidoreductase chain 4L	6	13	84	11.63	0.14
OPSD	*	Rhodopsin	4	7	351	149.83	0.43
PAHO	*	Pancreatic hormone	4	13	36	14.91	0.41
PCNA	*	Proliferating cell nuclear antigen	5	7	259	136.54	0.53
PRT1		Protamine B	4	10	16	4.38	0.27
PRVA	*	Parvalbumin α	4	10	100	44.12	0.44
RL2		50S Ribosomal protein L2	4	11	186	15.57	0.08
RL32		60S Ribosomal protein L32	4	8	39	3.10	0.08
RLA1	*	60S Ribosomal protein EL12′/EL12′-P	5	8	106	45.52	0.43
RLA2	*	80S Ribosomal protein EL12	4	7	111	50.23	0.45
RLUB		60S Ribosomal protein CEP52	6	9	51	33.14	0.65
RS12		30S Ribosomal protein S12	4	11	100	18.62	0.19
RS14		40S Ribosomal protein S14	5	11	94	16.90	0.18
RS19		30S Ribosomal protein S19	4	10	92	11.59	0.13
RS27	*	40S Ribosomal protein S27A	4	5	77	50.87	0.66
RS8		30S Ribosomal protein S8	4	8	128	13.86	0.11
SODC		Superoxide dismutase	8	20	146	51.84	0.36
SOMA		Somatotropin precursor	4	26	166	48.36	0.29
TBA1	*	Tubulin α-1 chain	7	12	441	311.43	0.71
TBA2		Tubulin α-2 chain	7	8	446	329.77	0.74
TBA3	*	Tubulin α-3 chain	4	4	444	350.96	0.79
TBB	*	Tubulin β chain	6	23	433	280.83	0.65
TBB1	*	Tubulin β-1 chain	6	10	443	316.71	0.71
TBB2	*	Tubulin β-2 chain	6	9	441	352.42	0.80
TOP2	*	DNA Topoisomerase II	4	5	1211	415.66	0.34
TPIS		Triosephosphate isomerase	6	11	245	102.24	0.42
TPMA	*	Tropomyosin α chain	4	6	284	265.41	0.93
TRYP	*	Trypsin I	4	7	222	84.26	0.38
UBIQ	*	Ubiquitin	6	12	76	71.96	0.95

Note. These are the gene families for which it was possible to find sequences in the SWISSPROT database, release 19, belonging to species from at least four of the taxonomic groups in Table 1. For each gene family, the table includes the number of taxa for which gene sequences were obtained; the total number of sequences in each species; the length of the maximal common pattern used to align those sequences; and the information content (IC) and information density (ID) of those patterns. To guarantee that the sequences are aligned in regions of unquestionable homology, only those genes were considered for which ID > 0.30 and those genes with ID > 0.25 and IC > 45. Averaged taxa distances were computed from the aligned regions of these genes, and the corresponding trees were obtained using the Fitch and Margoliash method. To identify topologically ill-defined trees, two additional tree reconstruction methods were used, maximal and minimal linkage. Only those genes for which maximal and minimal linkage resulted in the same clustering were further considered. The final 53 genes used in our analysis are identified with asterisks.

zoa constitute a single clade. It is interesting that in most cases in which fungi and metazoa are clustered together, protozoa taxa are represented by ciliate species, while in those cases in which metazoa are clustered with protozoa, the protozoa sequences often came from rhizopoda species.

In addition to such "direct" contradictions between individual gene trees, "indirect" inconsistencies may also arise when a global tree comprising all considered taxa is inferred from a set of noninclusive partial gene trees. For example, in the tree derived from histone H2A2, protozoa and embryophyta form a cluster, sister to the metazoa, while in the tree derived from histone H4, embryophyta and metazoa form a cluster, sister to

the chlorophyceae. Thus in the inferred combined tree, protozoa and embryophyta should constitute a cluster, sister to the chlorophyceae. However, such a clustering is never observed among our entire set of trees. Instead, in the clustering observed, derived from the sequences of some tubulins (TBA1, TBB1) and ubiquitin (UBIQ), chlorophyceae and protozoa belong to a cluster, sister to the embryophyta.

Some global contradictory results could be explained by the inclusion of a few very aberrant gene trees, perhaps constructed from poor alignments containing highly variable evolutionary rates. This seems unlikely, however, given our procedure. First, we considered only regions of statistically unquestionable similarity by using a local maximum similarity algorithm (Smith and Waterman, 1981). Second, we used a simple unweighted distance measure between the local regions. Third, we used only those genes for which maximum and single linkage clustering gave the same tree topology. The first of these makes it highly unlikely that we have aligned nonhomologous regions; the second minimizes the number of evolutionary assumptions employed; and the third ensures that the obtained tree topologies are robust to minor distance measure variation or errors.

Certainly there are still potential problems. Multiple substitutions in the same site mean that the observed number of amino acid differences is smaller than the actual number. This will result in an underestimation of the branch lengths, but does not necessarily affect the topology or branching order of the tree. On the other hand, we have implicitly assumed that the rates of amino acid substitution among different lineages are approximately the same. While the constant average rate hypothesis is controversial (Goodman, 1981; Wilson *et al.,* 1977; Czelusniak *et al.,* 1982; Wu and Li, 1985; Li *et al.,* 1987), it has been widely used in the estimation of divergence time and in the reconstruction of phylogenetic trees (Nei, 1975; Wilson *et al.,* 1977).

4. DISCREPANCIES BETWEEN GENE TREES CAN BE EXPLAINED BY GENE DUPLICATIONS

Contradictory gene trees are not necessarily wrong. It is well known that gene trees do not always reflect the true phylogenetic relationships between species, i.e., the species tree (Nei, 1987; Li and Graur, 1991). Gene duplication before divergence of distantly related species, followed by differential allelic losses after the splitting of more closely related species, can result in a gene tree in which distantly related species are branched together before the more closely related ones (Fig. 2A). Gene duplication has been suggested as one of the important mechanisms by which new genes can arise (Ohno, 1970). A duplicate gene would be free of many selective pressures, thus acquiring divergent mutations and eventually emerging as a new gene. The polymodal distribution of genome sizes, which has been registered in many groups of eukaryotes (Rees and Jones, 1972; Grime and Mowforth, 1982), strongly suggests that the entire genome of eukaryotes, most notably plants and higher animals, has undergone a number of near complete duplications.

Here we have developed a procedure for reconstructing the phylogenetic relationships among a number of taxa from a set of partial and contradictory gene trees. The reconstruction is obtained under the assumption that differences between gene trees are due to gene duplications and the consequent losses (either actual losses or failed recognition of duplicates). To compute the number of duplications and losses required to embed a gene tree into a species tree, we introduce in the next section the concept of mapping a gene tree into a species tree. Such a concept can be formalized within the context of graph theory. From such a formalization, computable (i.e., algorithmic) definitions of the (biological) concepts of duplication and loss can be obtained.

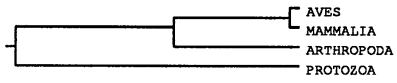
5. THE MAPPING OF A GENE TREE INTO A SPECIES TREE

Assume that a gene tree has been obtained for a number of species for which the species tree is known (Fig. 2A). A node in the gene tree can be assumed to correspond to a gene divergence event. The terminal leaves of the subtree generated by such a node correspond to the species carrying the alleles derived from such a gene divergence event. On the other hand, a node in the species tree corresponds to a species divergence (speciation) event. The terminal leaves of the subtree generated by such a node correspond to the species originated by such a speciation event. One can also think of a node from the species tree as corresponding to the most recent common ancestor of the species in the subtree generated by the node.

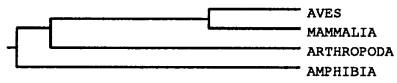
Gene divergence can be the result of either speciation or duplication. Thus nodes in a gene tree can corre-

FIG. 1. Trees obtained for 53 different gene families were constructed from matrices of averaged distances between taxa using the Fitch and Margoliash method. The root was placed at as near equal distance as possible from all terminal nodes. Branch lengths within a tree are approximate and are not comparable between different trees. Observe that trees obtained for different gene families are highly contradictory. Discrepancies between gene trees can be explained by gene duplications and the consequent losses during the course of evolution (see Fig. 2).

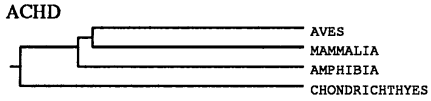
AACT



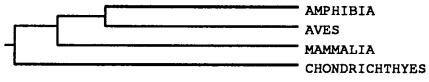
ACH2



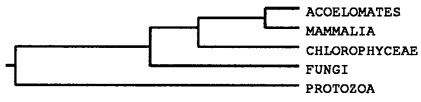
ACHD



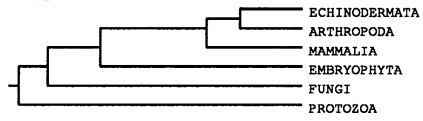
ACHG



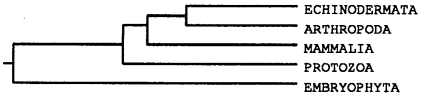
ACT



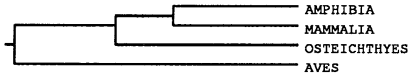
ACT2



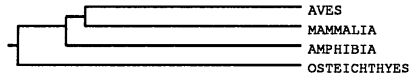
ACT3



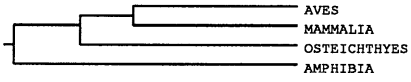
ACTB



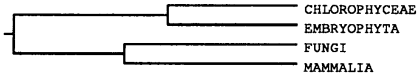
ANF



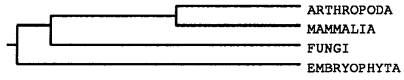
ANFC



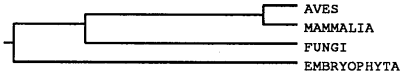
ATPB



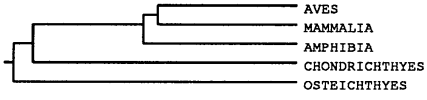
CATA



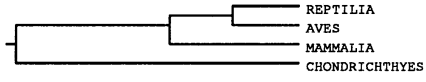
CISY



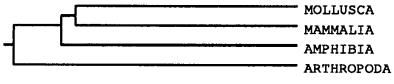
COLI



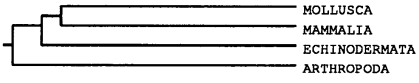
CRAA



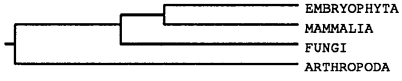
CYLA



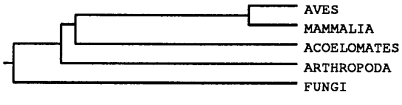
CYLB



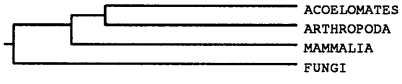
CYPH



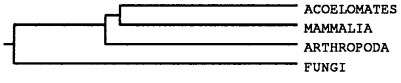
G3P



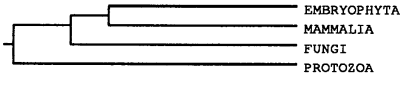
G3P1



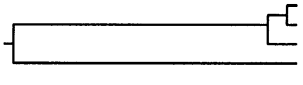
G3P2



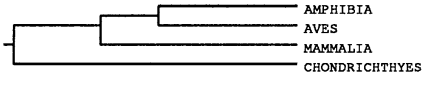
G6PI



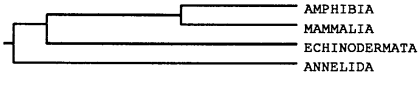
GAP1



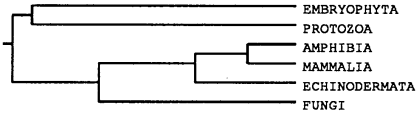
GLUC



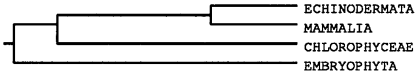
H1B



H2A2



H2A3



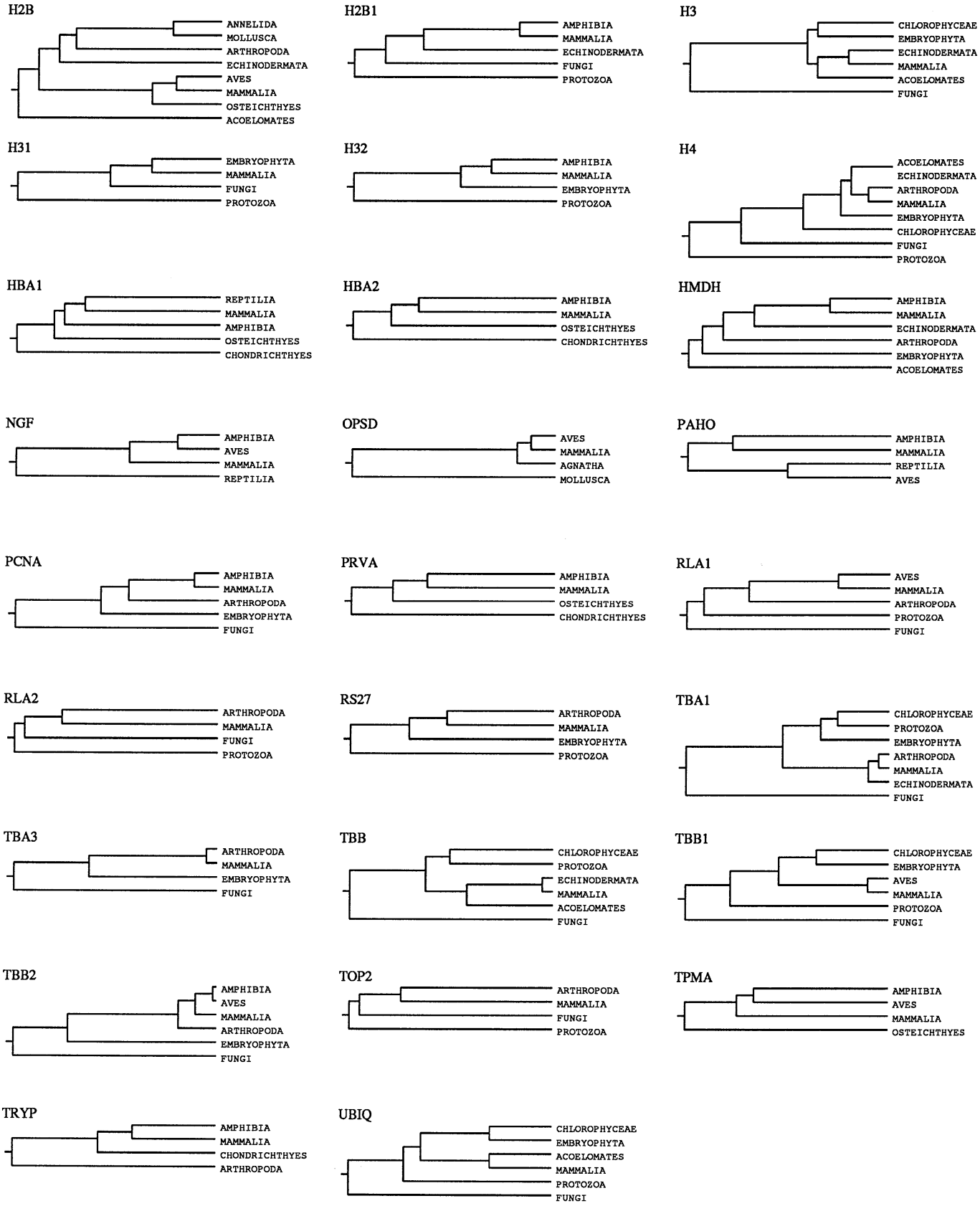


FIG. 1—*Continued*

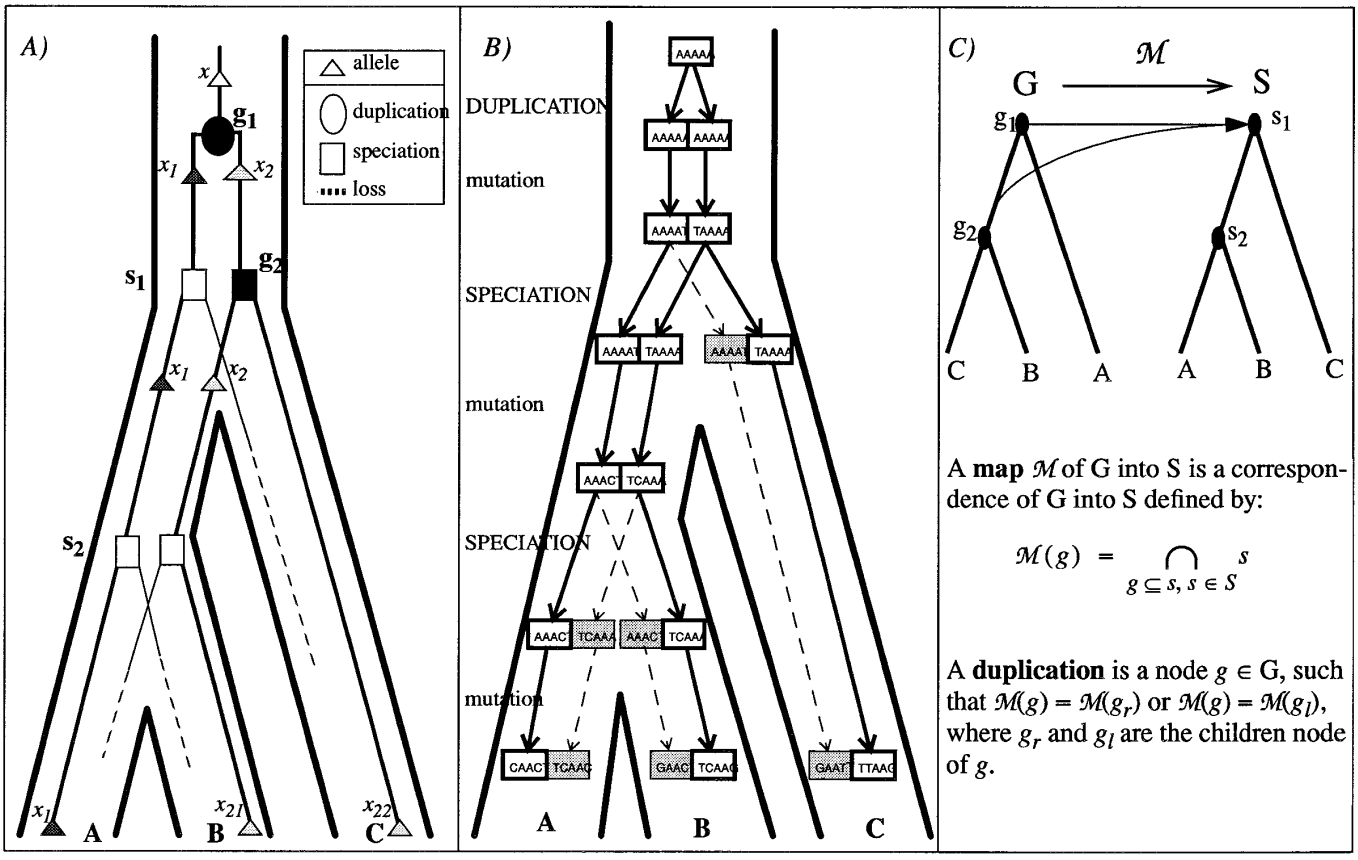


FIG. 2. A) A gene tree is obtained for a number of species for which the species tree is known. Gene trees do not necessarily reflect the true phylogenetic relationships between species. Thus the tree obtained from the sequence of the gene x—the gene tree T, its edges represented by intermediate width lines, and its nodes by black-filled symbols—does not reflect the true phylogenetic relationships between species A, B, and C—the species tree H, the pair of widest lines. Before C diverged from the lineage leading to A and B, the gene x duplicated. After the divergence, the two copies, x_1 and x_2, were initially preserved in the lineage leading to A and B, but x_1 was lost in the lineage leading to C. After A and B split, x_2 was lost in A, while x_1 was lost in B. Therefore C and B share a more closely related form of the ancestral gene x, than does A. Such is the relationship being reflected in the gene tree T. Genes x_1 and x_2 are said to be paralogous genes, since they have diverged from a duplication event, while x_{21} in B and x_{22} in C are orthologous genes, since they have diverged from a speciation event. (B) Using a "real" sequence to illustrate Fig. 2A. Because of gene duplication followed by differential allelic losses, species A and C, although more distant phylogenetically, share a more similar sequence than species A and B, which are phylogenetically closer. The shaded boxes correspond to lost, inactivated, or as yet unsequenced gene duplicates. (C) Gene duplication within graph theory. Nodes in the species tree S can be assumed to correspond to species divergence (speciation) events, while nodes in the gene tree G correspond to gene divergence events. We introduce the concept of the mapping of a gene tree into a species tree, which is a correspondence that maps gene divergence events into speciation events. Essentially, M maps a gene divergence event into the speciation event originating the species among which are found the actual species carrying the alleles derived from the gene divergence event. Thus the gene divergence event g_2 in G is mapped into the speciation event s_1 in S, since alleles x_{21} and x_{22}, originating from the gene divergence event g_2, are found in species B and C, which derive from the speciation event s_1. For the same reason, g_1 is mapped into s_1. Note that gene divergence can be the result either of speciation or of gene duplication. Therefore if a node in the gene tree corresponds to a speciation event (g_2 in G), the alleles deriving from it will be found among the species deriving from such a speciation event, and the node will be mapped into the node in the species tree corresponding to the same speciation event (s_1 in S). (Fig. 2A illustrates how g_2 corresponds to the speciation event s_1.) However if a node corresponds to a duplication event (g_1 in G), the alleles derived from it will be found among the species deriving from the next immediate speciation event, and it will be mapped into the node in the species tree corresponding to a speciation event (s_1 in S). In particular, a node in the gene tree corresponding to a duplication event will be mapped into the same node in the species tree as at least one of its children nodes, the one corresponding to next speciation event in the gene tree (g_2 in G). Thus in absence of gene duplication, allelic divergence occurs only with speciation, and different nodes in the gene tree are mapped into different nodes in the species tree. If gene duplication occurs, however, there is allelic divergence without speciation, and different nodes in the gene tree are mapped into the same node in the species tree.

The map M can be easily interpreted in A). The nodes g_1 and g_2 of the gene tree appear before s_1, and therefore they are both mapped into s_1.

spond either to speciation or duplication events. If gene divergence occurs only with speciation, gene and species tree are coincidental. However, if a gene is duplicated, gene divergence occurs without speciation and the gene phylogeny does not necessarily reflect the speciation history. For instance, if a duplication is carried across the next immediate speciation, and the two duplicated alleles are maintained in one of the lines of descent, but one of them is lost in the other, differential allelic losses after a subsequent speciation in the line of descent carrying the duplicated alleles can result in a gene tree with altered topology (see Figs. 2A and 2B). If a duplicated allele is lost before speciation, or if there are allelic losses immediately following the next immediate speciation, the duplication will not have an effect on the topology of the tree. In the first case, eventual gene divergence derived from the duplication will be lost before being carried across specific boundaries. In the second case, gene divergence derived from the duplication will be indiscernible from that derived from the speciation.

Given a near constant average mutation rate, there are no sources of gene divergence other than speciation and gene duplication. Thus discrepancies between gene and species trees will necessarily reflect the existence of gene duplications. The problem we now address can be stated in the following way: Given a gene tree for a number of species for which the species tree is known, determine the minimum number of gene duplications (i.e., only those affecting the topology of the tree), if any, that need to be postulated to conciliate gene and species trees. Our approach to this problem relies on distinguishing those nodes in the gene tree that correspond to gene duplication events from those nodes corresponding to speciation events. To distinguish nodes corresponding to duplication events, we introduce the concept of mapping a gene tree into a species tree. Essentially a node in the gene tree—corresponding to a gene divergence event—is assigned to the node in the species tree corresponding to the most immediate ancestor of the species carrying the alleles derived from a gene divergence event (see Fig. 2C).

The most immediate common ancestor of the species carrying the alleles derived from a duplication event will be exactly the species in which the duplication has occurred, the same most immediate common ancestor of the species carrying the alleles derived from the gene divergence event occurring with the next immediate speciation (see Figs. 2A and 2B). Thus if a node in the gene tree corresponds to a duplication event, it will be mapped into the same node in the species tree as at least one of its descendent nodes (children, grandchildren, great-grandchildren, . . . ,)—the one corresponding to the gene divergence event occurring with the next immediate speciation event. In particular, duplication nodes will be mapped into the same node as at least one of its children nodes, since nodes corresponding to multiple successive duplications before speciation will all be mapped into the same node, the one corresponding to the next immediate speciation event. Note that at least one of the descendent nodes of a duplication node will necessarily be a speciation node. The parent nodes of the terminal leaves of the gene tree are always speciation nodes, since there is never the need to postulate a duplication for two-leaf trees.

In what follows below, we formalize the concept of mapping a gene tree into a species tree and derive definitions for the biological concepts of *duplication* and *loss*. With such definitions, the number is computable for duplications and losses required to map a gene tree into a species tree. We use a local optimization algorithm to obtain the global tree minimizing the sum of the number of duplications and losses required to map into it, the set of partial gene trees obtained in Section 3 (Fig. 1).

Let G and S be binary trees with sets of leaves, L_G and L_S such that $L_G \subseteq L_S$. (In our biological interpretation, G would correspond to the gene tree and S to the species tree.) Let g be a node of G. g will also denote the set of leaves of the subtree of G whose root is g. Note that G is fully determined by its set of nodes, $G = \{g_1, \ldots, g_k\}$.

Definition. A map of G into S, $M: G \to S$ is the correspondence of G into S defined by

$$M(g) = \bigcap_{g \subseteq s,\, s \in S} s \text{ for all } g \in G.$$

Note that $M(g)$ is the smallest node in S, including g (the most immediate common ancestor of the species in the subtree generated by g, in our biological interpretation). Given a map M of G into S, we can define the concepts of *duplication* and number of *losses* associated to a node $g \in G$.

Definition. A duplication is a (nonterminal) node $g \in G$ such that

$$M(g) = M(g_r) \text{ or } M(g) = M(g_l),$$

where g_r and g_l are the children nodes of g.

That is, a duplication is a node in G mapped into the same node in S with at least one of its children nodes. Next, we define the concept of *number of losses* associated to a map. Let g_i and g_j be nodes of a binary tree G, such that $g_i \subseteq g_j$ and defines the distance between g_i and g_j, $d(g_i, g_j) = \text{car}\{g \in G: g_i \subset g \subseteq g_j\}$, where $car(x)$ is the number of elements in the set x. The distance $d(g_i, g_j)$ is the number of edges between nodes g_i and g_j. Let S_G be the *restriction* of S to G, that is, $S_G = \{x \cap L_G: x \in S\}$, and let M_G be the map of G into S_G.

Definition. Let g be a (nonterminal) node of G, the number of losses l_g associated to g is

$$l_g = \begin{cases} 0, & \text{if } M_G(g) = M_G(g_r) = M_G(g_l); \\ |d(M_G(g_r), M_G(g)) - 1| + |d(M_G(g_l), M_G(g)) - 1|, & \text{otherwise} \end{cases}$$

where $|x|$ is the absolute value of x.

The total number of losses to map G into S, l_G is $\sum_{g \in G} l_g$. If r_G is the total number of duplications, then $c(G, S) = l_G + r_G$ is the total number of duplications and losses required to map G into S.

6. OBTAINING THE OPTIMAL SPECIES TREE

Let S be a species tree built on a set of taxa L and let $G_1, \ldots, G_n$ be a collection of gene trees built on $L_1, \ldots, L_n$ where $L_i \subseteq L$ for all $i = 1, \ldots, n$, and $\bigcup_{i=1}^{n} L_i = L$.

(That is, the gene trees G_i are partial trees with respect to S since they do not necessarily comprise all taxa on which S is built.) The number $c(G_i, S)$ of duplications and losses required to map G_i into S can be computed by constructing the mapping M of G_i into S. Therefore the number

$$c(S) = \sum_{i=1}^{n} c(G_i, S)$$

of duplications and losses required to map $G_1, \ldots, G_n$ into S can be easily computed. In practice, however, S is unknown; only $G_1, \ldots, G_n$ are given, and the problem is to find S^* such that $c(S^*)$ is minimal. In principle, the optimal tree S could be found by exhaustively searching the set of all possible binary trees built on L. However, the number of these trees grows rapidly with car(L), and such an approach is computationally impractical. In Appendix 1 we describe a local minimization nearest neighbor branch swapping algorithm (Waterman and Smith, 1978) to find S^*. The algorithm produces a local optimum depending on an initial "seed" tree. Thus in Appendix 1, we also describe a preprocessing step intended to obtain a reasonably good initial species tree.

We have used the algorithm above to derive a global species phylogeny from the set of 53 gene trees obtained in Section 3. The obtainment of the initial species tree to be used as "seed" for the algorithm is described in Appendix 2. A total number of 91 duplications and 260 losses is required to map the 53 gene trees into the initial species tree. After applying the algorithm, we obtained a final species tree, shown in Fig. 4. The individual maps of the gene trees into the obtained species tree are shown in Fig. 3. A total number of 46 duplications and 101 losses is required to map the 53 partial gene trees into a species tree.

Initial and final species trees differ in several important aspects. First, whereas in the initial tree the nonmetazoan taxa (fungi, protozoa, chlorophyceae, and embryophyta) formed a single clade, in the final tree, the chlorophyceae–embryophyta lineage is a sister group of the metazoan lineage. Second, the invertebrate tree has been radically rearranged: echinodermata has been placed closer to chordata than to the remaining invertebrates, and the acoelomate and mollusca lineages are no longer constituting a single clade, but mollusca and annelida form a sister group of the arthropoda lineage, and acoelomates appear as the earliest group within protostomes.

It is important to mention that for only 18 of the gene trees (about one-third) there is no need to postulate duplications (Fig. 3), that is, most of the trees are inconsistent with the postulated phylogeny—even though such a phylogeny is the globally more consistent with the individual gene trees. This suggests that only a fraction of the molecular phylogenies may reflect the true species phylogeny and, therefore, that the usage of molecular sequence data to infer phylogenetic histories may suffer from severe limitations (see Discussion).

7. GENOME DUPLICATIONS

Gene duplications do not need to be independent events. Duplications of different genes may be the result of a large-scale genomic duplication. The various gene duplications we have identified among our set of gene trees may therefore be traced back to only a few large-scale genome duplications. From the maximum parsimony viewpoint this requires the determination of the minimum number of (complete) genome duplications required to explain the individual gene duplications observed and their location in the species phylogeny.

The determination of this number relies essentially on the assumption that gene duplications occurring at the same location in the species phylogeny are the result of the same genome duplication. For instance, we will assume that the duplications occurring at the ACHG and GLUC gene trees (Fig. 3) result from the same genome duplication, occurring before the separation of amphibia from the remaining tetrapoda. Note, however, that we often have considerable freedom in placing gene duplications into the species tree. Such is the case, for instance, in the duplication occurring at the G3P gene tree (Fig. 3). The duplication has occurred before the splitting of deuterostomes and protostomes, but after the separation of fungi. Therefore the duplication could be placed before the splitting of deuterostomes and protostomes, or before the splitting of meta-

zoans and plants, or before the separation of protozoa, but not before the separation of fungi. In general, a gene duplication is assumed to have occurred anywhere between the node in the species phylogeny in which it is mapped and the node in which its parent is mapped—between the node in which it is mapped and the root of the species phylogeny in the particular case in which the duplication occurs at the root of the gene tree.

We make use of this freedom. Again following a parsimonious principle, we cluster the gene duplications into the minimum number of locations in the species tree, with the constraint that all gene duplications clustered into a given location are allowed to occur in such a location. In Appendix 3, we outline an heuristic procedure to solve the problem. Gene duplications clustered into the same location in the species tree could be the result of more than one genome duplication (since multiple genome duplications may have occurred in the time expanse between two consecutive species tree nodes). We cluster all gene duplications occurring at the same location into the minimum required events.

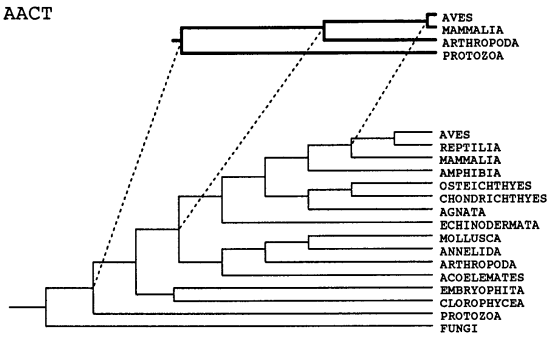
Only five independent genome duplications at four different locations must be postulated in order to explain the 46 individual gene duplications observed. Their locations are shown in Fig. 4. Also identified are the individual gene duplications associated with each of the proposed genome duplications. Three locked duplications (i.e., duplications that cannot be placed anywhere else, see Appendix 3) support a genomic duplication occurring before the separation of amphibia from the remaining tetrapoda. Two locked duplications support the genomic duplication occurring before the splitting of protostomes and deuterostomes. One double locked gene duplication supports genomic duplication occurring before the separation of protozoa. Finally, 13 gene duplications have been assigned to the root of the species phylogeny, supporting a minimum of one additional genomic duplication.

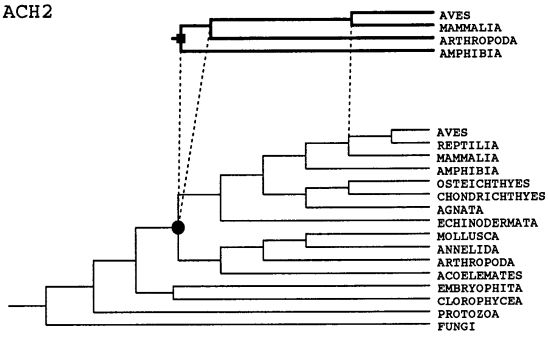
8. DISCUSSION

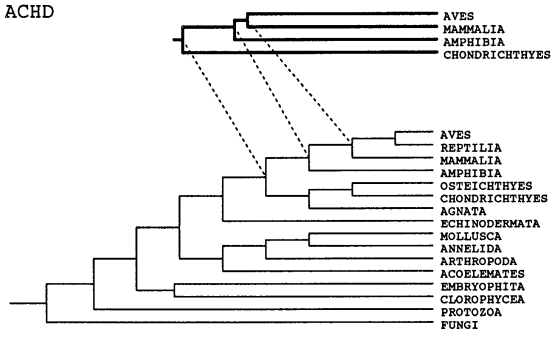
We have investigated a highly contradictory set of trees among distantly related eukaryotic groups obtained using sequences from different gene families. As genome projects progress, larger sets of gene sequences from distantly related organisms will become available. This vast amount of data will undoubtedly be very valuable in reconstructing the phylogenetic history of living beings; however, it is likely to be highly contradictory, as we have observed. Contradictions will arise even though one has carefully followed procedures to minimize the errors in inferring the trees from the sequence data. Phylogenetic reconstruction methods need to be developed in order to handle intrinsically inconsistent data. We have developed and tested a method to reconstruct a species phylogenetic tree from a large set of inconsistent gene trees under the assumption that differences among gene trees are the consequence of gene duplications. The method finds the global phylogeny minimizing the number of duplications and losses that need to be postulated. The results have been presented largely as a proof of concept.

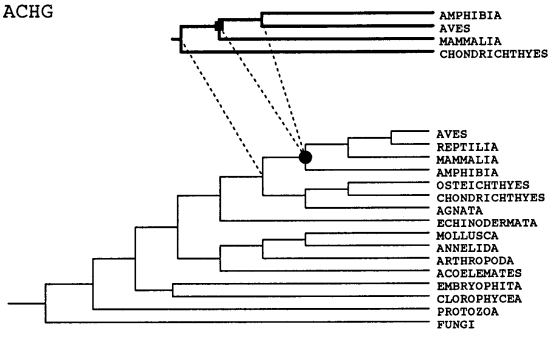
We have used the method to reconstruct the phylogenetic relationship in Fig. 4 among 16 major higher eukaryotic taxa from a set of 53 different genes. Obviously, we do not claim to have solved the eukaryotic phylogeny, but the tree that we have obtained appears very reasonable. Its main features are the following: Fungi appear to be the earliest group among higher eukaryotes. Although such a hypothesis is reasonable from a morphological perspective, it has been recently suggested from a comparison of small subunit ribosomal RNA sequences, that the fungal and metazoan lineages share a more recent common ancestor than either does with plants (Wainright *et al.*, 1993). The position of protozoa, sister to fungi, is not very relevant, since we have included within protozoa, ciliate, rhizopoda and other species that may occupy different phylogenetic positions. Our phylogeny supports the monophyletic origin of the relationship between land plants and green algae. Phylogenetic relationships among metazoa, on the other hand, are extremely controversial (Morris, 1993). Different phylogenies have been derived from exactly the same molecular sequence data (Field *et al.*, 1988; Lake, 1990). In that context, the phylogeny we present is reasonable. Although metazoa appear as a monophyletic group in such a phylogeny, we cannot really address such a controversial issue since cnidaria, which have been suggested to arise from a protist ancestry different from bilateria (Field *et al.*, 1988), are missing from the taxa we have considered. Within bilateria, our placement of acoelomates—linked to protostomes—is possibly the most controversial and contradicts previous phylogenies (Field *et al.*, 1988). The clear phylogenetic separation, however, between deuterostomes and protostomes is well-recognized (Morris, 1993), as well as the consideration

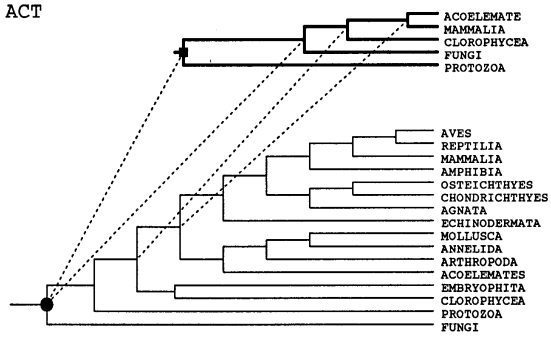
FIG. 3. The maps of each of the individual 53 gene trees into the consensus or species phylogeny. Bolded trees correspond to the gene trees. The thinner trees correspond to the species phylogeny. Dashed lines show the mapping of each gene tree into the species phylogeny. Multiple nodes in the gene tree mapped into the same species tree node correspond to gene duplication events. They are identified by a black-filled square in the gene tree. Their mapping in the species phylogeny indicates the location in the species phylogeny where the duplication occurred (identified by a black-filled circle).

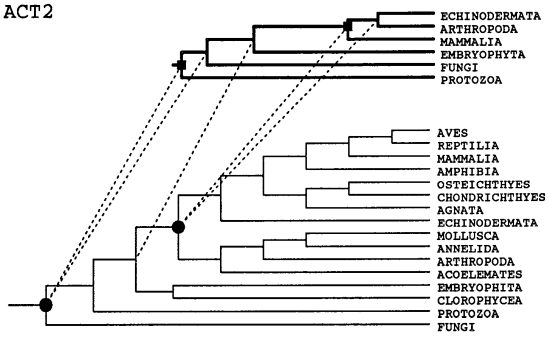

AACT
AVES
MAMMALIA
ARTHROPODA
PROTOZOA
AVES
REPTILIA
MAMMALIA
AMPHIBIA
OSTEICHTHYES
CHONDRICHTHYES
AGNATA
ECHINODERMATA
MOLLUSCA
ANNELIDA
ARTHROPODA
ACOELEMATES
EMBRYOPHITA
CLOROPHYCEA
PROTOZOA
FUNGI

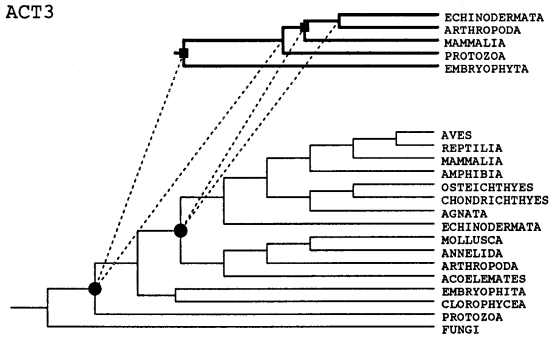

ACH2
AVES
MAMMALIA
ARTHROPODA
AMPHIBIA
AVES
REPTILIA
MAMMALIA
AMPHIBIA
OSTEICHTHYES
CHONDRICHTHYES
AGNATA
ECHINODERMATA
MOLLUSCA
ANNELIDA
ARTHROPODA
ACOELEMATES
EMBRYOPHITA
CLOROPHYCEA
PROTOZOA
FUNGI

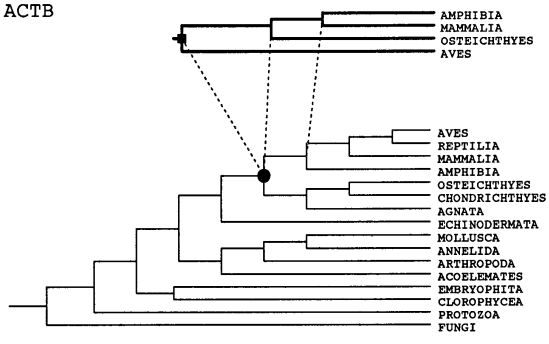

ACHD
AVES
MAMMALIA
AMPHIBIA
CHONDRICHTHYES
AVES
REPTILIA
MAMMALIA
AMPHIBIA
OSTEICHTHYES
CHONDRICHTHYES
AGNATA
ECHINODERMATA
MOLLUSCA
ANNELIDA
ARTHROPODA
ACOELEMATES
EMBRYOPHITA
CLOROPHYCEA
PROTOZOA
FUNGI


ACHG
AMPHIBIA
AVES
MAMMALIA
CHONDRICHTHYES
AVES
REPTILIA
MAMMALIA
AMPHIBIA
OSTEICHTHYES
CHONDRICHTHYES
AGNATA
ECHINODERMATA
MOLLUSCA
ANNELIDA
ARTHROPODA
ACOELEMATES
EMBRYOPHITA
CLOROPHYCEA
PROTOZOA
FUNGI


ACT
ACOELEMATE
MAMMALIA
CLOROPHYCEA
FUNGI
PROTOZOA
AVES
REPTILIA
MAMMALIA
AMPHIBIA
OSTEICHTHYES
CHONDRICHTHYES
AGNATA
ECHINODERMATA
MOLLUSCA
ANNELIDA
ARTHROPODA
ACOELEMATES
EMBRYOPHITA
CLOROPHYCEA
PROTOZOA
FUNGI


ACT2
ECHINODERMATA
ARTHROPODA
MAMMALIA
EMBRYOPHYTA
FUNGI
PROTOZOA
AVES
REPTILIA
MAMMALIA
AMPHIBIA
OSTEICHTHYES
CHONDRICHTHYES
AGNATA
ECHINODERMATA
MOLLUSCA
ANNELIDA
ARTHROPODA
ACOELEMATES
EMBRYOPHITA
CLOROPHYCEA
PROTOZOA
FUNGI


ACT3
ECHINODERMATA
ARTHROPODA
MAMMALIA
PROTOZOA
EMBRYOPHYTA
AVES
REPTILIA
MAMMALIA
AMPHIBIA
OSTEICHTHYES
CHONDRICHTHYES
AGNATA
ECHINODERMATA
MOLLUSCA
ANNELIDA
ARTHROPODA
ACOELEMATES
EMBRYOPHITA
CLOROPHYCEA
PROTOZOA
FUNGI


ACTB
AMPHIBIA
MAMMALIA
OSTEICHTHYES
AVES
AVES
REPTILIA
MAMMALIA
AMPHIBIA
OSTEICHTHYES
CHONDRICHTHYES
AGNATA
ECHINODERMATA
MOLLUSCA
ANNELIDA
ARTHROPODA
ACOELEMATES
EMBRYOPHITA
CLOROPHYCEA
PROTOZOA
FUNGI

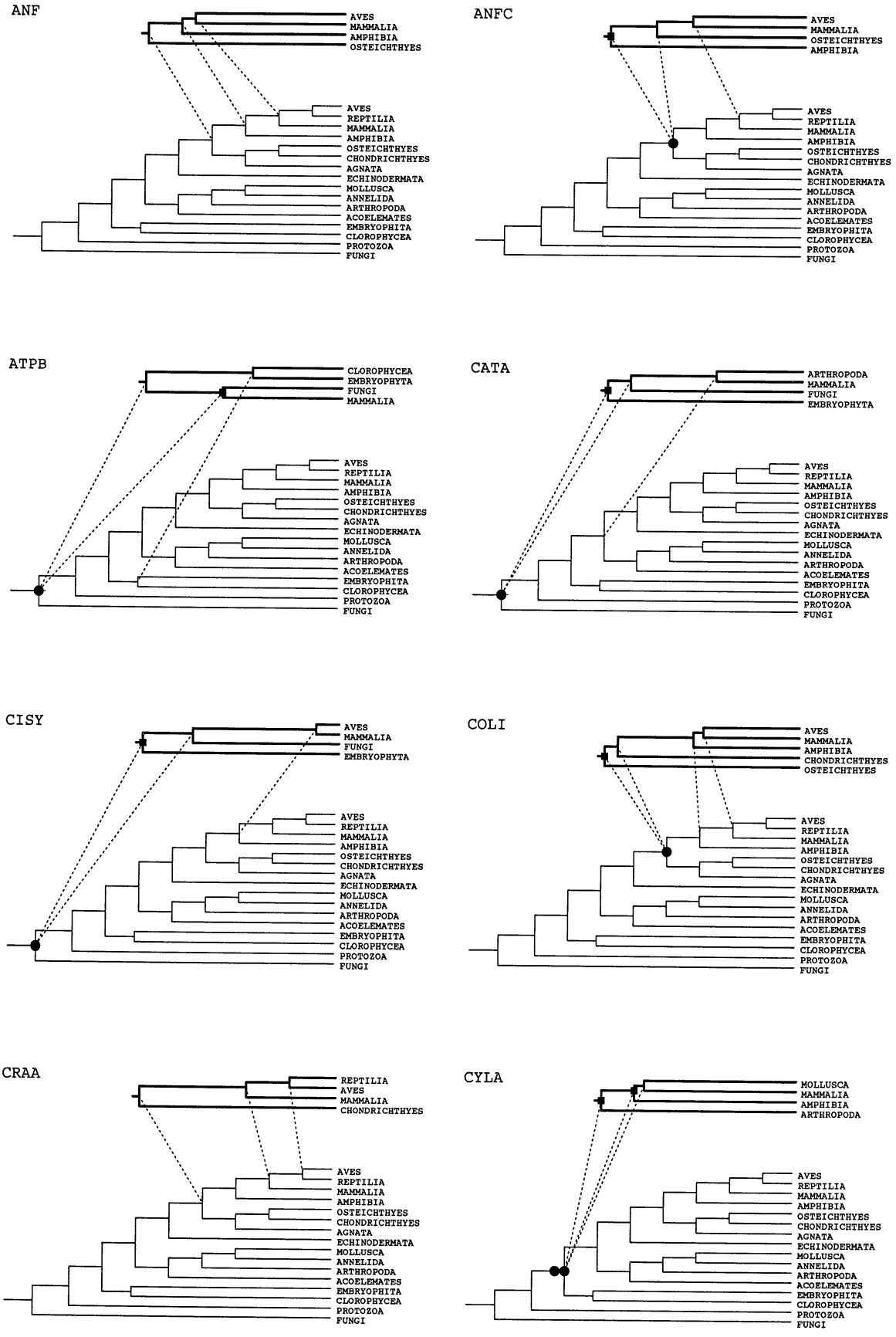


FIG. 3—*Continued*

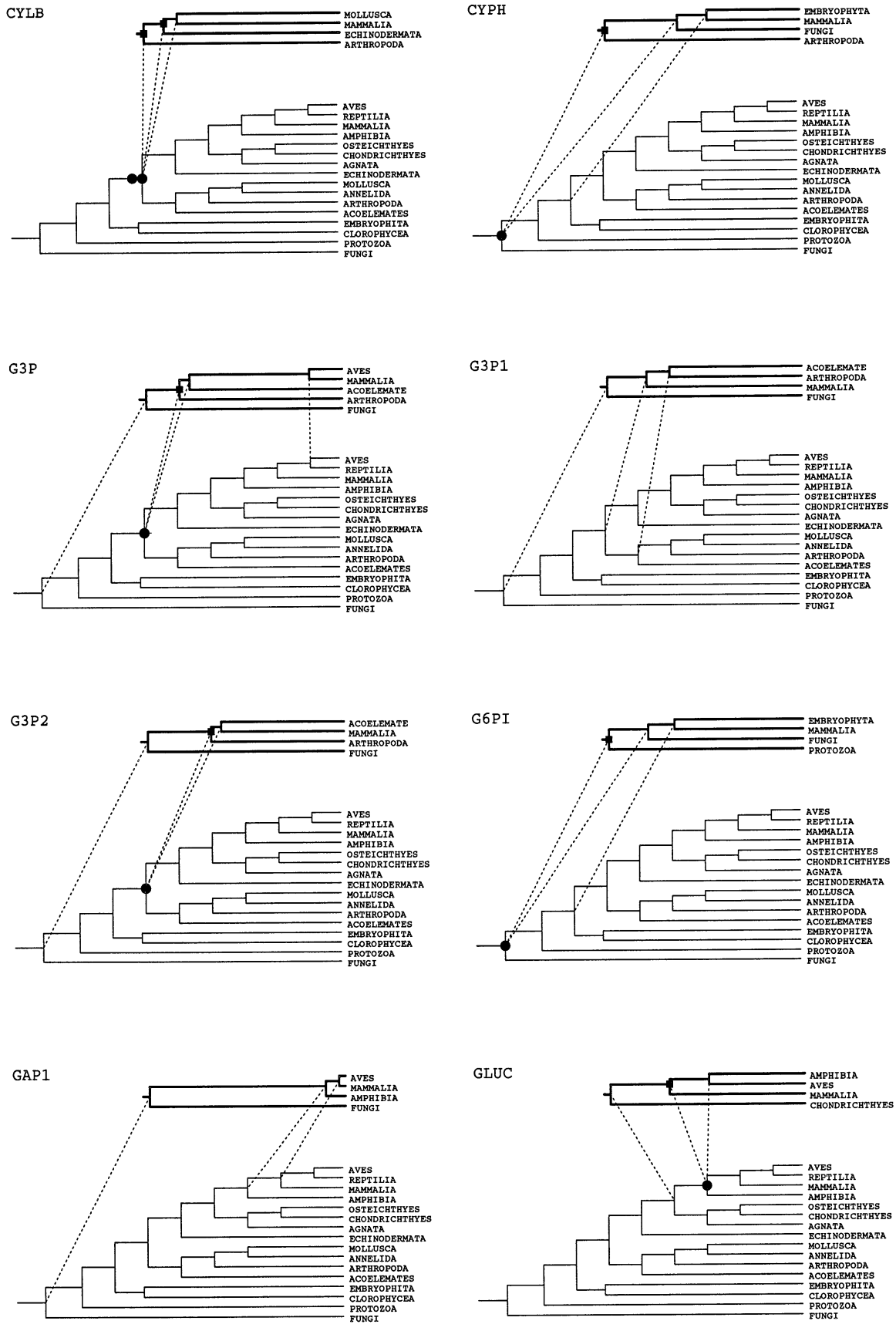


FIG. 3—*Continued*

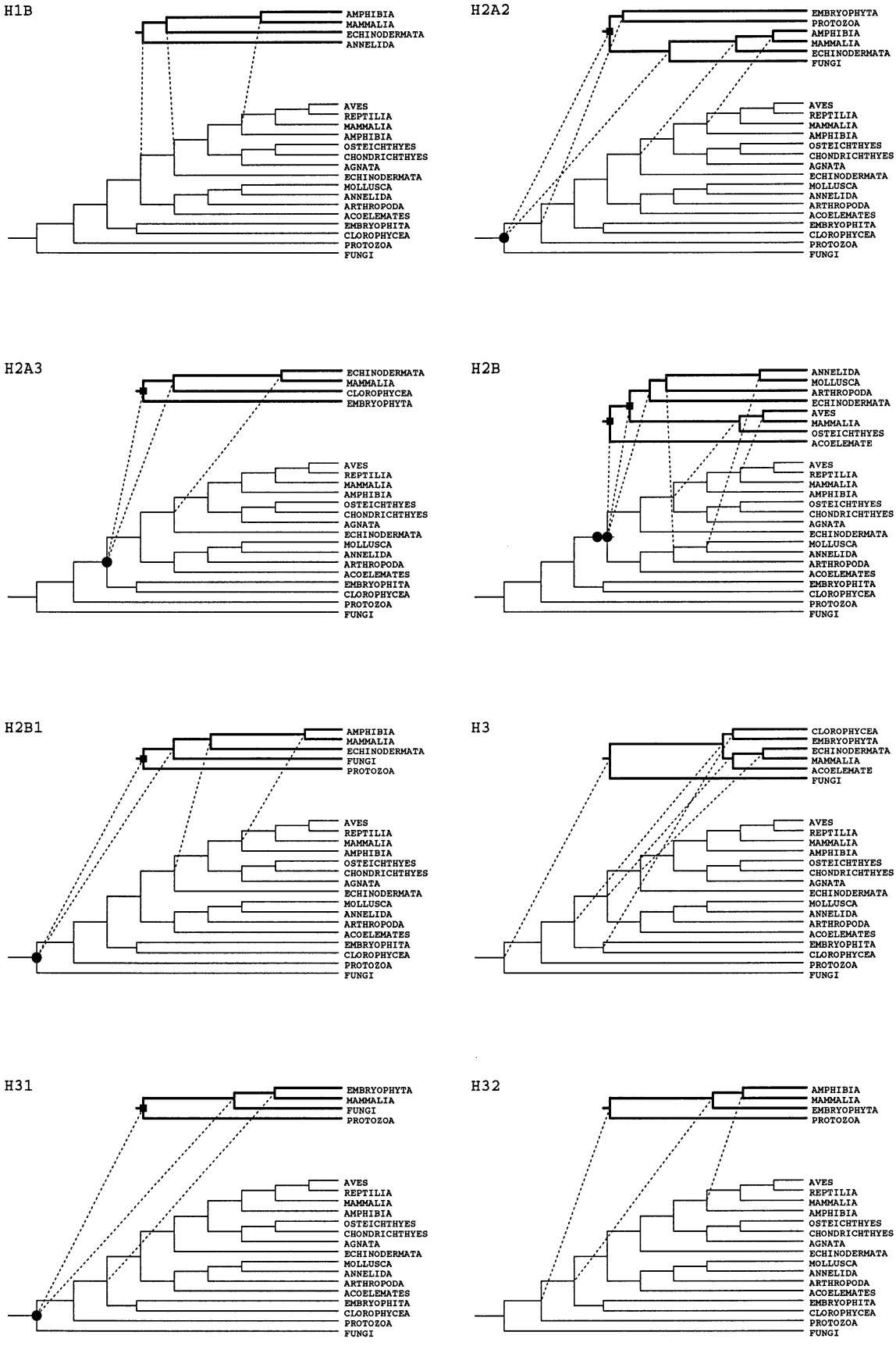


FIG. 3—*Continued*

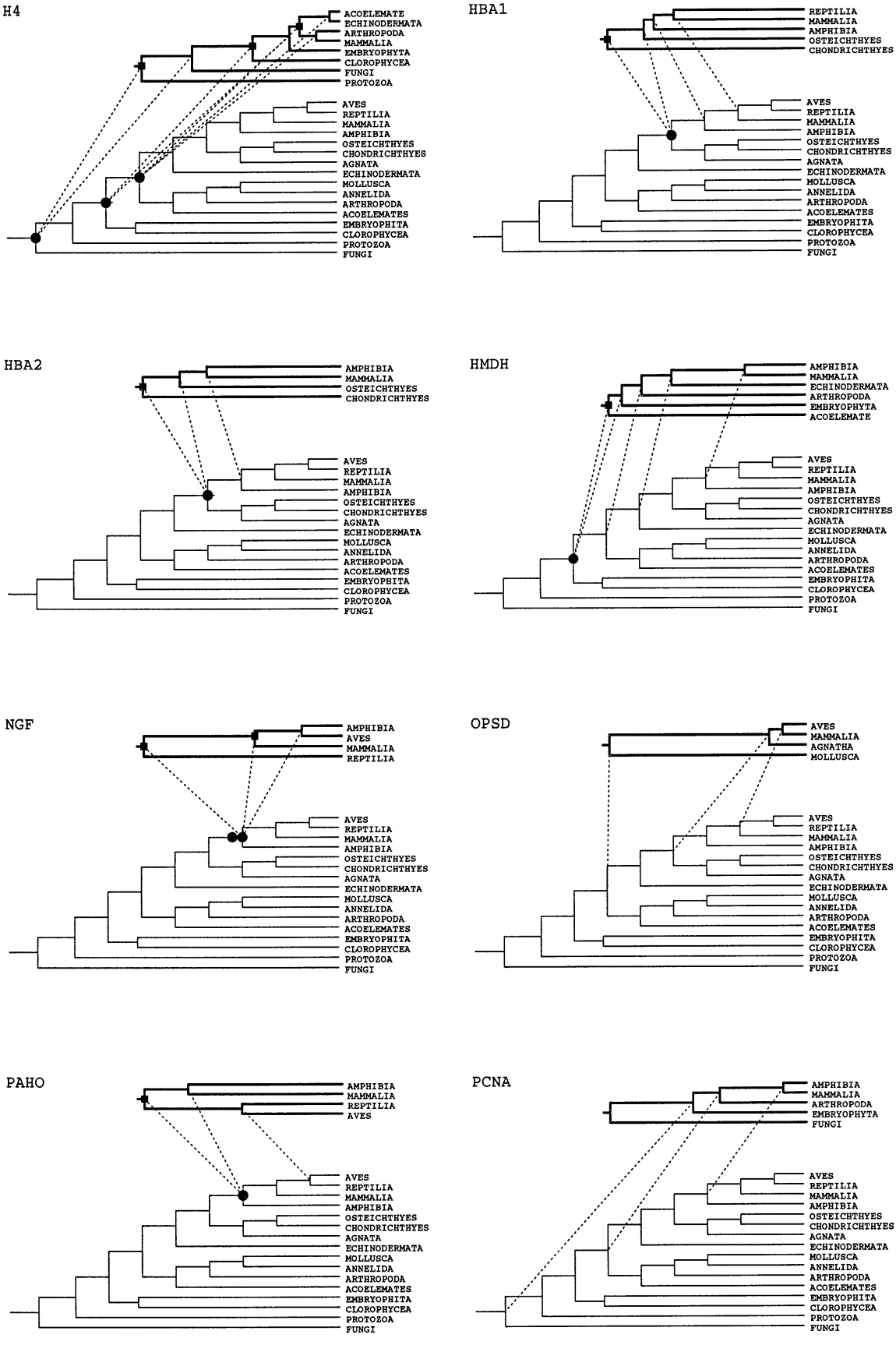


FIG. 3—*Continued*

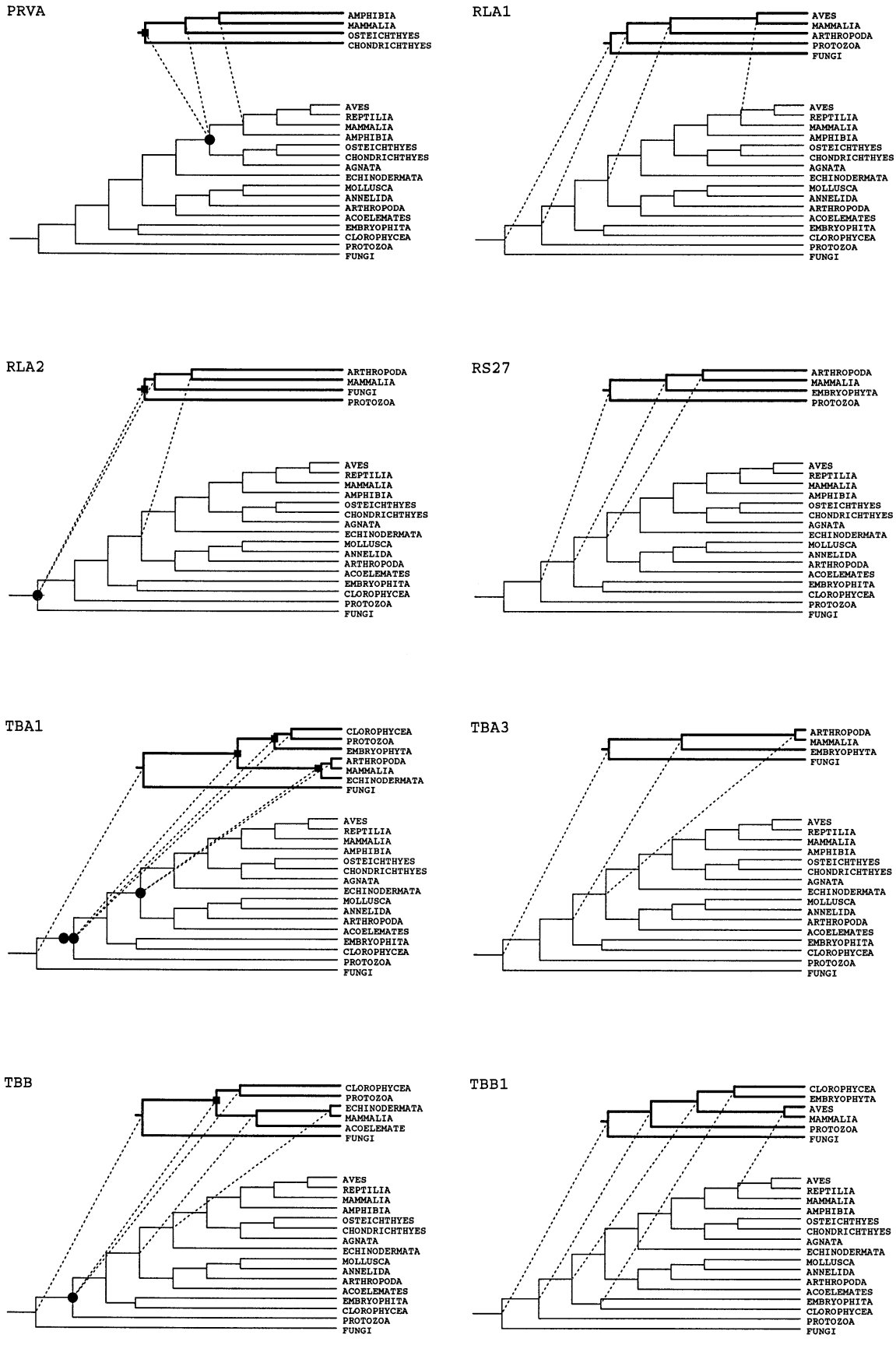


FIG. 3—*Continued*

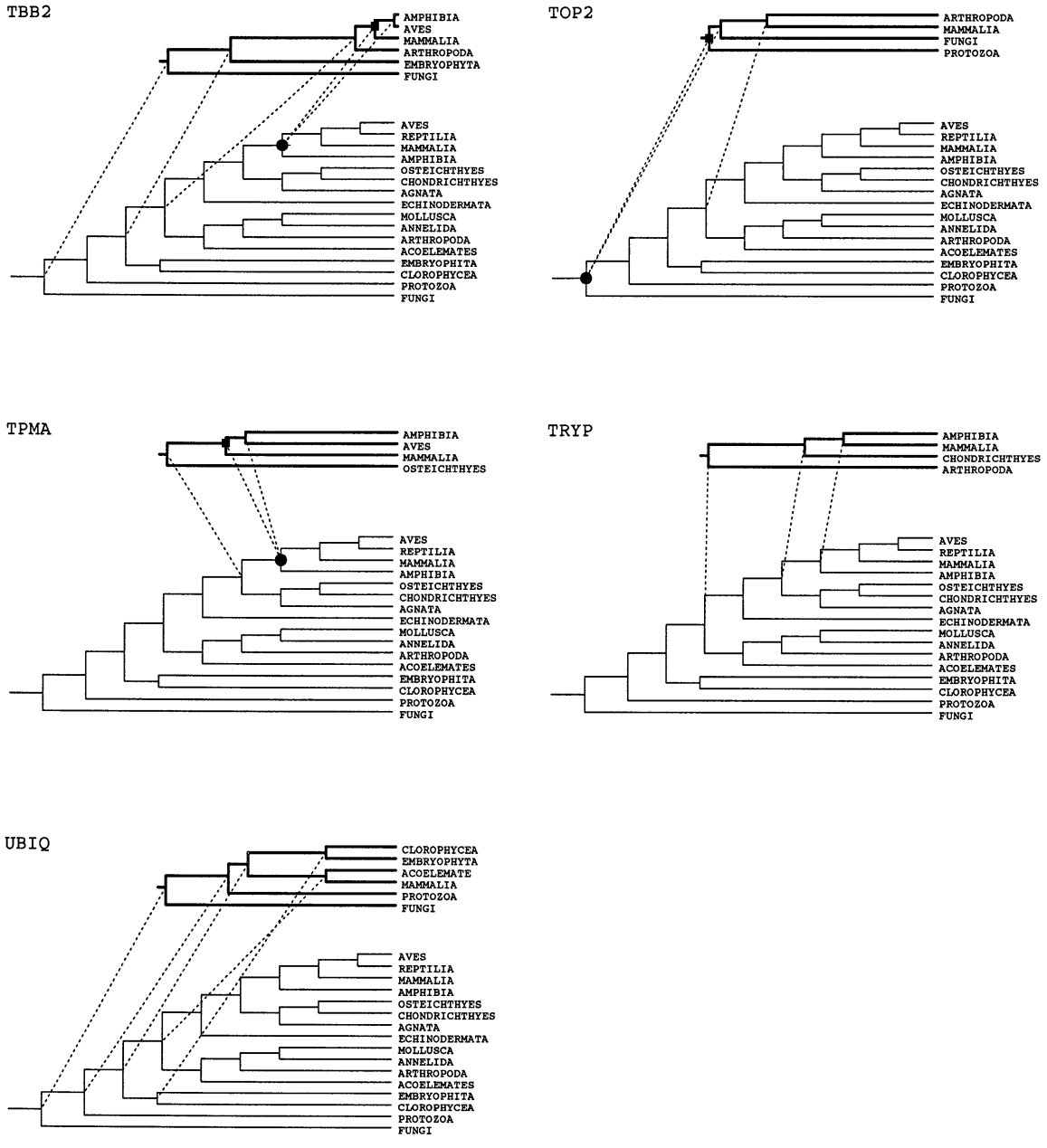


FIG. 3—*Continued*

within protostomes of the anellida–mollusca lineage as a sister group of the arthropoda lineage (Lake, 1990) and the monophyletic origin of deuterostomes (Field, 1988). Within chordata, our phylogeny supports the fishes as constituting a single clade sister to the tetrapoda. Such a relationship, however, is incompatible with a possible monophyletic origin for gnathostomes (Forey and Janvier, 1993). Our tree cannot address the controversial issue of the monophyletic origin of the cyclostomes (Stock and Whitt, 1992), because all agnatha have been included in a single clade. Similarly, we can not investigate the polyphyletic origin of tetrapoda because all reptiles have been included in the same clade, but phylogenetic relationships within tetrapoda appear reasonable.

Only 18 among the 53 individual gene trees are consistent with the above eukaryotic phylogeny. This has important implications for attempted phylogenies constructed from only one or two genes. For the remaining 35 gene families, a total of 46 gene duplications needs

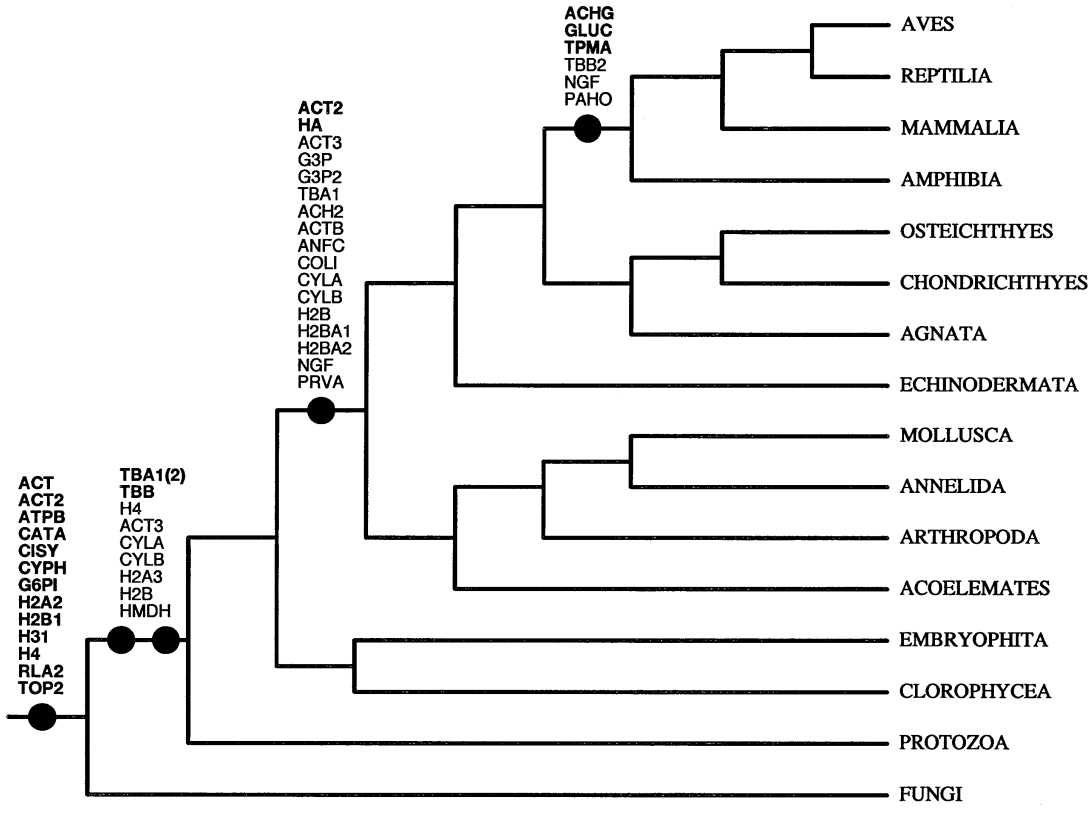


FIG. 4. The consensus eukaryotic species phylogeny is obtained. Black circles show the location of the minimum number of genome duplications required. We have assumed that gene duplications occurring at the same location in the species phylogeny are most probably the result of the same single large scale genome duplication event. Only five genome duplications at four different locations need to be postulated to explain the 46 gene duplications observed. At each genome duplication, we list the individual gene duplications that would be associated with it. Bold typeface denotes absolutely locked duplications (see Section 7 and Appendix 3).

to be assumed to conciliate the gene trees with the inferred phylogeny. That such a large percentage of trees appears to be inferred from paralogous genes may be surprising. We have already pointed out, however, that the genome of higher eukaryotes has probably undergone a number of near complete duplications, and in consequence, that most eukaryotic genes have probably been duplicated (Section 4). Distinguishing the paralogous from the truly orthologous genes may be a difficult task, in particular if orthologous genes have evolved to a different function, while paralogous genes have evolved to the same, or related, function. In addition, at the current stage of the sequencing of the genome of the living organisms, it cannot be guaranteed that the truly orthologous genes from distantly related species have already been sequenced and included in the databases. Examples exist in which genes thought to be orthologous have later been discovered to be actually paralogous, once the truly orthologous gene has been sequenced (Sumi *et al.*, 1992; Forterre *et al.*, 1993). Whether most of the cases of paralogy we have identified are purely artifactual—the result of the truly orthologous gene not yet being sequenced—or real—the result of the orthologous genes being differentially lost or evolving to different functions—is an issue we do not think can be settled at this point. We expect that as genome sequencing progresses, numerous cases of artifactual paralogy will be recognized, but it is possible for a substantial number of genes to be intrinsically paralogous across the phylogenetic spectrum.

We have related individual gene duplications to a few global genome duplications. We have developed a method to cluster and map individual gene duplications in the gene trees into global genome duplications in the species tree (Section 7). In the particular case we have studied, the 46 individual gene duplications can be shown to be compatible with only five genome duplications occurring at four different locations across eukaryotic history (Fig. 4). The fact that a substantial fraction of the gene duplications observed are *locked* (see Appendix 3) and that they tend to cluster in a few locations (see Figs. 3 and 4) offers some guarantee that our placement of the genome duplications is not completely artificial. It appears reasonable. Interestingly, the fact that two independent genome duplications separate fungi from the remaining higher eukaryote lin-

eages correlates with the fact that the size of the fungal genome is substantially smaller than that of the other higher eukaryotes (Cavalier-Smith, 1985) and further strengthens the hypothesis that fungi represent the oldest higher eukaryotic lineage.

The work presented here should in no way be considered complete. Rather, a number of questions, regarding both the quality of the data and the methodology employed, remain open for further research. The completeness and correctness of the data set could be improved by using a larger sample of genes, considering particularly those genes for which sequences exist for a large number of species across the phylogenetic spectra. The presence of artifactual paralogy in the data set should also be carefully investigated. Methodological improvements can be made at both the inference of individual gene trees and the reconstruction of the global phylogeny. Higher quality gene trees may be obtained by computing more accurate distance matrices between sequences, taking into account, for example, uneven rates of evolutionary change among different groups, or by using likelihood methods to infer the gene trees from the sequence data.

Similarly, improvements can be considered in the construction of the global phylogeny. The correctness of the nearest neighbor branch swapping local minimization algorithm should be investigated, for example, by determining how sensible the final solution is to the phylogenetic tree taken as initial seed. On the other hand, weights could be assigned to the individual gene trees to correct for homologous gene families' unequal representation in the data set or to take into account the adequacy of the individual gene tree's topology and branch lengths to the original distances in the distance matrix.

The issues of the reliability and robustness of the global phylogeny should also be addressed. The very same relative number of duplications and losses required for a global phylogeny already provides an indication of its reliability. For instance, a global phylogeny obtained from a small number of mostly topologically inconsistent gene trees seems in principle less reliable than a phylogeny derived from a large number of topologically consistent gene trees. The robustness of the phylogeny, on the other hand, could be studied by determining the neighborhood of the best phylogeny obtained. If the next best phylogeny requires a similar number of duplications and losses, and it is topologically very similar to the best phylogeny, evidence for such a phylogeny is weaker than if the next best phylogeny requires a substantially larger number of duplications and has a similar topology. In the case of a weak phylogeny, the addition of new gene trees may radically alter the resulting topology.

Finally, once the topology of the global phylogeny has been established, the problem of determining its branch lengths (i.e., the divergence times between the taxa) could be addressed. Since branch lengths for the global phylogeny are inferred from branch lengths of individual trees obtained for genes with very different rates of evolutionary change, both the branch lengths of the global phylogeny and the rate of evolutionary change of the individual genes should be estimated simultaneously.

We do not claim to have obtained the ultimate eukaryotic phylogeny nor the precise location where genome duplications have occurred during eukaryotic history. This was not our goal. Rather it was to develop a conceptual framework, reasonable from an evolutionary standpoint, in which apparently contradictory molecular genetic data can be coherently integrated. With the increasing pace at which gene families are being sequenced in distantly related organisms, we expect contradictory sequence data to arise with increasing frequency. The integration of contradictory data will be essential in order to solve any phylogeny, as well the integration of other morphological, embryological, and paleontological data. We hope that the work presented here will contribute to such an integration.

APPENDIX 1

Let S be a species tree built on a set of taxa L and let $G_1, \ldots, G_n$ be a collection of gene trees built on $L_1, \ldots, L_n$ where $L_i \subseteq L$ for all $i = 1, \ldots, n$, and $\bigcup_{i=1}^{n} L_i = L$.

(That is, the gene trees G_i are partial trees with respect to S since they do not necessarily comprise all taxa on which S is built.) The number $c(G_i, S)$ of duplications and losses required to map G_i into S can be computed by constructing the mapping M of G_i into S. Therefore, the number $c(S) = \sum_{i=1}^{n} c(G_i, S)$ of duplications and losses required to map $G_1, \ldots, G_n$ into S can be easily computed. In practice, however, S is unknown; only $G_1, \ldots, G_n$ are given, and the problem is to find S^* such that $c(S^*)$ is minimal. In principle, the optimal tree S could be found by exhaustively searching the set of all possible binary trees built on L. However, the number of these trees grows rapidly with car (L), and such an approach is computationally impractical.

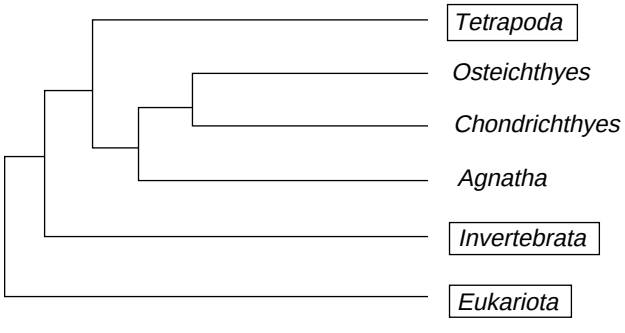
Here we use a local minimization algorithm to find an optimal tree S^*. This algorithm proceeds iteratively from an initial seed tree S_1. At a given iteration t, the neighborhood of the seed tree for such an iteration, S_{1t}, is searched. The neighborhood of S_{1t}, $N(S_{1t})$, is obtained by nearest neighbor branch swapping (Waterman and Smith, 1978). For each tree $S \in N(S_{1t})$, the number $c_G(S^*)$ is computed. The best tree, S_{1t}^*, found in $N(S_{1t})$, is taken as a seed tree for the next iteration ($S_{1t+1} = S_{1t}^*$). The procedure is repeated until an iteration T is reached such that the neighborhood of the seed tree for the iteration does not contain a better tree

($S^*_{IT} = S_{IT}$). In such a case, the seed tree is a local optimum, S^*_I. Note that the procedure does not guarantee that the optimum found, S^*_I, is the global one, S^*.

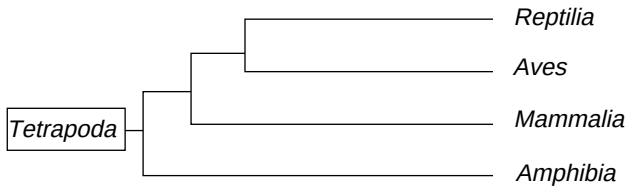
Since the local optimum obtained by above algorithm, S^*_I, may be very dependent on the initial seed tree used, S_I, we have designed a preprocessing step intended to obtain a reasonably good initial species tree. In such a preprocessing step, the taxa considered in L are first aggregated into a few reasonable supertaxa, and a simplified biologically reasonable tree relating such supertaxa is constructed. Second, the supertaxa are sequentially desegregated. Such desegregation is obtained in the following way: the terminal node corresponding to the supertaxa in the simplified tree is substituted by each possible subtree within the supertaxa; each of the resulting trees is used as a different initial tree, and a local minimum (i.e., minimizing the total number of duplications and losses) for the simplified tree given the set of gene trees in G obtained in each case. The supertaxa are finally substituted by the subtree found in the local minimum the greatest number of times. Third, once all supertaxa have been desegregated and substituted by the optimal subtree, the resulting tree is used as the initial seed tree for the local minimization algorithm to search the complete tree.

APPENDIX 2

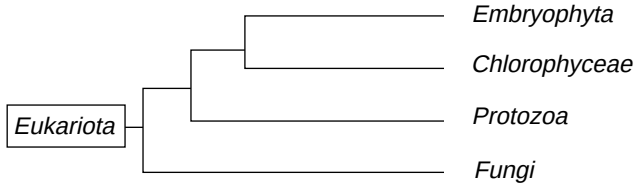
Here we describe the obtainment of the initial seed species tree given the 16 taxa that we consider and the 53 genes obtained in Section 3. We proceeded in the following way. We created the supertaxa tetrapoda, including amphibia, aves, mammalia, and reptilia; the supertaxa invertebrata, including acoelomates, annelida, arthropoda, echinodermata, and mollusca; and the supertaxa eukariota, including chlorophyceae, embryophyta, fungi, and protozoa. We built the following simplified tree,



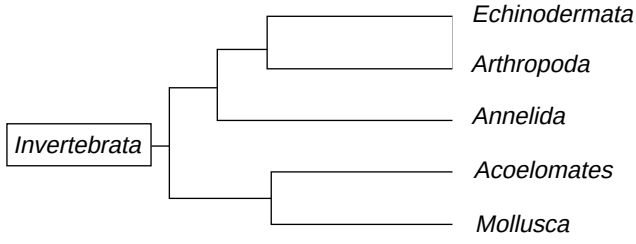
We first desegregated the tetrapoda supertaxa. Fifteen different subtrees can be obtained within the tetrapoda supertaxa. Each one of them was replaced in above simplified tree, and the local minimum for such a tree was obtained. The following subtree for the tetrapoda supertaxa was found in eight of the minima:



Consequently, the tetrapoda supertaxa were substituted by such a subtree in the above simplified tree. Next, the eukariota supertaxa were desegregated. Similarly, 15 subtrees can be obtained within eukariota, and each of them was used in a different initial tree. In 12 of the minima obtained, the eukariota subtree was the following:



The eukariota supertaxa was substituted by such a subtree. Finally, the invertebrata supertaxa were desegregated. Mollusca and acoelomates were joined together in a single taxon, and 15 different subtrees were built within invertebrata. In this case, however, 3 different minima were obtained the same number of times. One of them was arbitrarily chosen to substitute the invertebrata supertaxa:



The tree resulting from the successive desegregations of tetrapoda, eukariota, and invertebrata was used as the initial tree for the local minimization algorithm to search the complete tree.

APPENDIX 3

Here we formally describe the problem of clustering individual gene duplications into a few genomic duplications and placing such duplications in a species tree. In our formalization, we first define the set of allowed positions for a given gene duplication to occur in the species phylogeny. Then we define a class of functions, which we term *gmaps*, that map a given set of gene duplications into a node in the species tree, with the restriction that all gene duplications in the set are al-

lowed to occur in the node in which the set is mapped. Then, being that D is the set of duplications needed to map a set of gene trees $G_1, \ldots, G_n$ into a species tree S, we state the problem we are trying to solve in the following terms: Find the minimum cardinality partition of D for which there exists a *gmap* into S. At the end of the Appendix we provide an heuristic algorithm to find such a partition. At this point, we have a clustering of gene duplications into a few locations in the species phylogeny. However, multiple genome duplications can occur at a single location where a set of gene duplications has been mapped. We therefore need to assign gene duplications occurring at a single location to a number of genomic duplications (possibly only one). We proceed in the following way. We define the *order* of a gene duplication d in a cluster of gene duplications as the number of duplications in the cluster included in d. Finally, gene duplications having the same order are assigned to the same genomic duplication.

In what follows, $p(g)$ denotes the parent of an internal node g from a binary tree G. And if x is a pair $x = (x_1, x_2)$, then $\Pi_1(x) = x_1$ and $\Pi_2(x) = x_2$ (that is, Π_1 and Π_2 are the first and second projections of x.)

3.1 Allowed Locations for Gene Duplications in Species Tree

Let M be a map of G into S, and let $g \in G$ be a duplication.

Definition. The set of allowed locations for g in S, A_g is

$$A_g = \begin{cases} \{s \in S: M(g) \subseteq s \subseteq M(p(g))\} & \text{if } g \text{ is an internal or terminal node} \\ \{s \in S: M(g) \subseteq s\} & \text{otherwise.} \end{cases}$$

A_g is the set of nodes in the species tree S located between the map of a duplication and the map of the parent of such duplication.

Let $G_1, \ldots, G_n$ be binary trees built on $L_1, \ldots, L_n$, and let S be a binary tree built on

$$L = \bigcup_{i=1}^{n} L_i.$$

Definition. Let D be the set defined in the following way

$$D = \{(x, i): x \in G_i \text{ is a duplication}, i = 1, \ldots, n\}.$$

D is the set of duplications required to map $G_1, \ldots, G_n$ into S. Since duplications from different trees can be indistinguishable, we identify the tree to which a duplication belongs.

3.2 Gmaps

Let $P_D = \{Q_D^1, \ldots, Q_D^k\}$ be a partition of D. (That is, P_D is a clustering of the set of duplications obtained when mapping $G_1, \ldots, G_n$ into S.)

Definition. A function $F: P_D \to S$ is a *gmap* if and only if for all $Q_D^i \in P_D$:

$$\text{if } g \in Q_D^i, \text{ then } F(Q_D^i) \in A_g.$$

A function $F: P_D \to S$ maps sets of gene duplications into nodes in S. If F is a *gmap*, then a set of gene duplications, $Q_D^i \in P_D$, is mapped into a node in S, $F(Q_D^i)$, such that all gene duplications in Q_D^i, $g \in Q_D^i$, are allowed in $F(Q_D^i)$—that is, $F(Q_D^i)$ belongs to the set of allowed positions of all gene duplications in Q_D^i.

Definition. A partition P_D of D is a *gmap partition* of D given S, if there exists a *gmap* of P_D into S.

The problem. Given D, the set of duplications obtained when mapping $G_1 \ldots, G_n$ into S, obtain G_D^*, a minimum cardinality *gmap partition* of D given S. The problem is thus to find the clustering of gene duplications into the smallest number of clusters such that each cluster is mapped into a node in S in which all gene duplications in the cluster are allowed to occur. G_D^* provides the clustering of the gene duplications, while a *gmap* $F: P_D \to S$ provides the nodes in S where the cluster of gene duplications are postulated to occur. We have designed an heuristic procedure to find simultaneously both G_D^* and $F: P_D \to S$. The procedure is described at the end of this Appendix.

3.3 Genome Duplications

Gene duplications within a given cluster are further clustered into *genome duplications.* Note that we cannot assume that all gene duplications in a given cluster correspond to the same genome duplication, because multiple genome duplications can occur at the same location. To obtain the genome duplications that need to be postulated at each location and the set of gene duplications than can be traced back to each genome duplication, we proceed in the following way.

Let G_D^* be a minimum cardinality *gmap partition* of D, and let $H \in G_D^*$ be a cluster of gene duplications.

Definition. For each $g \in H$, we define the *order* of g, d_g

$$d_g = \text{car}\left\{ \begin{matrix} h \in H, \text{ such that } \Pi_2(h) = \Pi_2(g) \\ \text{and } \Pi_1(h) \subseteq \Pi_1(g) \end{matrix} \right\}.$$

For a given duplication g in a cluster H, d_g is simply the number of duplications in the cluster that are contained in g. That is, d_g is the order of occurrence of g in a case of multiple duplications. Obviously, if g does not occur in a multiple duplication, $d_g = 1$.

Let R_H be the relation defined in H by: "$g\ R_H\ h$ if and only if $d_g = d_h$", and let

$$GD_H = G_D^*/R_H$$

be the partition induced by R_H in G_D^*. Each member of GD_H is a set of gene duplications assigned to the same genomic duplication (and we will indeed say that $x \in GD_H$ is a *genome duplication*.)

$$GD = \bigcup_{H \in G_D^*} GD_H$$

is the minimum set of genome duplications that need to be postulated to map $G_1, \ldots, G_n$ into S.

3.4 An Algorithm to Find G_D^*

Let $G^1, \ldots, G_n$ be gene trees and S and species tree. Let D be the set of duplications needed to map $G_1, \ldots, G_n$ into S. We define three types of postulated duplications:

(1) A duplication g is *free* if it occurs at the root of the corresponding gene tree.
(2) A duplication g is *locked* if it occurs at an interval node in the gene tree.
(3) A locked duplication g is *absolutely locked* if the map of the parent node of the duplication is the parent node of the map of the duplication. Note that in such a case the duplication can be placed at only one location ($\mathrm{car}(A_g) = 1$).

We then proceed in the following way. First, we place in the global phylogeny the duplications that are absolutely locked. Second, we place the duplications that are locked, but not absolutely. Those locked duplications that can occur at a location where an absolutely locked duplication occurs are placed at such a location. In the case in which a locked duplication can occur in more than one location where an absolutely locked duplication occurs, the locked duplication is placed at the closest location preceding the node in the species phylogeny in which the duplication is mapped—that is, the location furthest from the root of the species phylogeny. For example, the locked duplication occurring at the G3P gene tree can be placed at two locations where absolutely locked duplications occur: before the splitting of deuterostomes and protostomes, or before the separation of protozoa (see Fig. 3). The duplication is placed at the first of these two locations, because such a location is the closest one preceding the node in the species phylogeny in which the G3P gene duplication is mapped. For those locked duplications that cannot occur at locations where absolutely locked duplications occur, the set of locations (defined by nodes in the species phylogeny) where they can occur is determined. A duplication whose set of potentially occurring locations does not overlap the set of potentially occurring locations for another locked duplication is placed at the closest location preceding the node in which the duplication is mapped. If the sets of potentially occurring locations for different locked duplications overlap, the locations are chosen such that the number of duplications in the phylogeny is minimized. (Usually the number of overlapping sets and their size will be small enough, such that the locations can be found by simple inspection of such sets. Incidentally, in the case we are studying here, we have not found locked duplications that cannot occur at locations where absolutely locked duplications occur.) Finally, free duplications are placed at the closest location preceding the node in which the duplication is mapped where a duplication—absolutely locked or locked—, if any, has already been placed. Otherwise, they are placed at the root of the phylogeny. After this process, we obtain both G_D^* and a *gmap* of G_D^* into S.

ACKNOWLEDGMENTS

We thank Walter Fitch, Morris Goodman, and Richard Holmquist for valuable discussions and helpful criticisms. R.G. did part of this work during a workshop in the Aspen Center for Physics. This work was supported in part by Grant P41 LM05205-12 from the National Library of Medicine. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the granting agency.

REFERENCES

Cavalier-Smith, T. (1985). Selfish DNA and the origin of introns. *Nature* **315:** 283–284.

Czelusniak, J., Goodman, M., Hewitt-Emmett, D., Weiss, M. L., Venta, P. J., and Tashian, R. E. (1982). Phylogenetic origins and adaptive evolution of avian and mammalian haemoglobin genes. *Nature* **298:** 297–300.

Field, K. G., Olsen, G. J., Lane, D. J., Giovanni, S. J., Ghiselin, M. T., Raff, E. C., Pace, N. R., and Raff, S. J. (1988). Molecular phylogeny of the animal kingdom. *Science* **239:** 748–753.

Fitch, W. M., and Margoliash, E. (1967). Construction of phylogenetic trees. A method based on mutation distances as estimated from cytochrome c sequences is of general applicability. *Science* **155:** 279–284.

Forey, P., and Janvier, P. (1993). Agnathans and the origin of jawed vertebrates. *Nature* **361:** 129–134.

Forterre, P., Benachenbou-Lafha, N., and Labedan, B. (1993). Universal tree of life. *Nature* **362:** 29.

Goodman, M. (1961). Decoding the pattern of protein evolution. *Prog. Biophys. Mol. Biol.* **38:** 105–164.

Goodman, M., Czelusniak, J. G., Moore, W., Romero-Herrera, A. E., and Matsuda, G. (1979). Fitting the gene lineage into the species lineage: a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.* **28:** 132–168.

Grime, J. P., and Mowforth, M. A. (1982). Variation in genome size and ecological interpretation. *Nature* **299:** 151–153.

Hardisty, M. W. (1979). "Biology of Cyclostomes," Chapman & Hall, London.

Hasegawa, M., and Hashimoto, T. (1993). Ribosomal RNA trees misleading? *Nature* **361:** 23.

Lake, J. A. (1990). Origin of the metazoa. *Proc. Natl. Acad. Sci. USA* **87:** 763–766.

Li, W.-H., and Graur, D. (1991). "Fundamentals of Molecular Evolution," Sinauer, Sunderland, MA.

Li, W.-H., Tanimura, M., and Sharp, P. M. (1987). An evaluation of the molecular clock hypothesis using mammalian DNA sequences. *J. Mol. Evol.* **25:** 330–342.

Morris, S. C. (1993). The fossil record and the early evolution of the Metazoa. *Nature* **361:** 219–225.

Nei, M. (1975). "Molecular Population Genetics and Evolution," North-Holland, Amsterdam.

Nei, M. (1987). "Molecular Evolutionary Genetics," Columbia Univ. Press, New York.

Ohno, S. (1970). "Evolution by Gene Duplication," Springer-Verlag, Berlin.

Page, R. D. M. (1994). Maps between trees and cladistic analysis of historical associations among genes, organisms and areas. *Syst. Biol.* **43:** 58–77.

Rees, H., and Jones, R. N. (1972). The origin of the wide species variation in nuclear DNA content. *Int. Rev. Cytol.* **32:** 53–92.

Smith, R. F., and Smith, T. F. (1992). Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling. *Protein Engineering* **5:** 35–41.

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.

Sogin, M. L., Hinkle, G., and Lelpe, D. D. (1993). Universal tree of life. *Nature* **362:** 29.

Stock, D. W., and Whitt, G. S. (1992). Evolutionary implications of the cDNA sequence of the single lactate dehydrogenase of a lamprey. *Proc. Natl. Acad. Sci. USA* **89:** 1799–1803.

Sumi, M., Sato, M. H., Denda, K., Date, T., and Yoshida, M. (1992). A DNA fragment homologous to F_1-ATPase β subunit was amplified from genomic DNA of *Methanosarcina barkeri*. *FEBS Lett.* **3:** 207–210.

Wainright, P. O., Hinkle, G., Sogin, M. L., and Stickel, S. K. (1993). Monophyletic origins of the Metazoa: An evolutionary link with fungi. *Science* **260:** 340–342.

Waterman, M. S., and Smith, T. F. (1978). On the similarity of dendrograms. *J. Theor. Biol.* **73:** 789–800.

Wilson, A. C., Carlson, S. S., and White, T. J. (1977). Biochemical evolution. *Biochemistry* **46:** 573–639.

Wu, C.-I., and Li, W.-H. (1985). Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl. Acad. Sci. USA* **82:** 1741–1745.