MOLECULAR PHYLOGENETICS AND EVOLUTION
Vol. 1, No. 3, September, pp. 242–252, 1992

Split Decomposition: A New and Useful Approach to Phylogenetic Analysis of Distance Data

Hans-Jürgen Bandelt* and Andreas W. M. Dress†

*Mathematisches Seminar, Universität Hamburg, D-2000 Hamburg 13, Federal Republic of Germany; and †Fakultät für Mathematik, Universität Bielefeld, D-4800 Bielefeld 1, Federal Republic of Germany

Received December 12, 1991; revised October 7, 1992

In order to analyze the structure inherent to a matrix of dissimilarities (such as evolutionary distances) we propose to use a new technique called split decomposition. This method accurately dissects the given dissimilarity measure as a sum of elementary "split" metrics plus a (small) residue. The split summands identify related groups which are susceptible to further interpretation when casted against the available biological information. Reanalysis of previously published ribosomal RNA data sets using split decomposition illustrate the potential of this approach. © 1992 Academic Press, Inc.

INTRODUCTION

Phylogenetic analysis of molecular sequence data often is carried out by first calculating pairwise similarity coefficients, converting these into evolutionary distances, and finally applying some distance-matrix method in order to estimate an unrooted phylogenetic tree. Goodness-of-fit would be judged by comparing the evolutionary distances with the additive distances read off the estimated tree. So, data are fit to a best (or at least, near-optimal) tree, whether or not they bear any resemblance with additive tree data. In practice, one tries to avoid methodological artifacts by applying different tree approximation methods (some operating on sequence data, others using derived distances) and then putting up with a strict consensus tree. Still, one may fall into the trap of systematic error when the methods are subject to the same bias and all disguise true phylogenetic relationships.

We therefore propose to accompany any phylogenetic analysis by a nonapproximative method as well that allows for conflicting alternative groupings (to some extent) and hence is able to detect some of those distinctive minor features in distance data which are dominated by others and not supported by estimated trees. This goal can be achieved by *split decomposition*, developed by Bandelt and Dress (1992), which may be regarded as a kind of factor analysis for distance matrices. It decomposes any dissimilarity matrix d into a number of "binary factors," described as "splits" weighted by "isolation indices," plus a residual indecomposable term (here interpreted as noise). For phylogenetic analysis split decomposition serves two purposes: (a) to exhibit tentative phylogenetic relationships even when they are overridden by parallel events, and (b) to detect groupings brought about by pronounced convergence or systematic error.

As to point (a), assume a phyletic line separates two pairs of taxa 1, 2 and 3, 4; then with respect to phyletic distance p the sum $p_{12} + p_{34}$ (of distances between 1 and 2, 3 and 4) is smaller than $p_{13} + p_{24} = p_{14} + p_{23}$ ("additivity" alias "4-point condition"). Evolutionary distances d, presenting only an estimate of true phyletic relationships, normally do not even respect the ordering $d_{12} + d_{34} < d_{13} + d_{24}$ and $d_{12} + d_{34} < d_{14} + d_{23}$, but one could hope that at least $d_{12} + d_{34}$ is not the largest of the three sums.

Given this as a working hypothesis, we would then expect such a pattern to be observed whenever the two taxa 1 and 2 are chosen from a group $\mathcal{J}$ which is separated from its complementary group $\mathcal{K}$ by a phyletic line, while taxa 3 and 4 are chosen from the complement $\mathcal{K}$. Consequently, any complementary pair $\mathcal{J}$, $\mathcal{K}$ satisfying this (comparatively weak) condition will be called a d-split.

To any such d-split one can, moreover, associate a positive weight, the *isolation index* (see Eq. (2) below), which in the case of additive distances would yield the length of the corresponding branch in the representing tree. However, there may be more d-splits than those supported by true phylogenetic relationships. These, typically exhibiting a low isolation index, often reflect traits of penetrating parallelism.

To illustrate point (b), imagine that an observed distance matrix d is the sum $d = p + e$ of a matrix p of linearly scaled phyletic distances plus an error term e such that e itself happens to be realized by some tree different from the one representing p. Then the d-splits would consist exactly of all splits which are either p-

1055-7903/92 $5.00

splits or e-splits or both (with isolation indices of p and e adding up to the indices for d). Which of the splits belong to p and which to e, though, cannot be decided unambiguously. If the error term e has considerably smaller entries than p, then the d-splits with larger isolation indices would belong to p rather than e.

The theory of split decomposition predicts at most $\binom{n}{2}$ d-splits for any n by n distance matrix (Bandelt and Dress, 1992; Theorem 3, p. 60, and Corollary 4, p. 62). This bound is considerably larger than $2n - 3$, the maximum number of splits in a tree connecting n taxa, yet it is small enough to have all d-splits computed efficiently. Reanalysis of numerous distance matrices derived from sets of n aligned ribosomal RNA sequences (with n between 10 and 25, say) confirms that biologically relevant data typically bring about $2n$ splits, a large portion of which fit together on a single tree, and leave a small residue. In contrast, randomly generated distance matrices tend to have a rather large residue and to produce mostly *trivial* splits, separating one taxon from all the remaining ones, and only very few others, generally separating no more than two or at most three taxa from the rest.

In this paper we present a few illustrative case studies which we have chosen more or less arbitrarily from the existing literature and which we believe to be somewhat representative.

METHODS

Split Decomposition

Assume we are given a matrix $d = (d_{ij})$ of dissimilarities between pairs of taxa $1, \dots, n$. For any four taxa i, j, k, l we compare the three distance sums $d_{ij} + d_{kl}$, $d_{ik} + d_{jl}$, $d_{il} + d_{jk}$. If i, j, k, l were located on a tree such that there is a link separating i, j from k, l, then the sum $d_{ij} + d_{kl}$ (with respect to the additive path length metric d) would be the smallest among those three sums. This pattern would thus be shown by any two pairs i, j and k, l separated by a fixed link of the tree, so that this link and its length can be reconstructed from the associated distance matrix. Since real data are far from such an ideal tree situation, we relax the criterion for accepting a partition of the taxa into two parts $\mathfrak{J}$, $\mathfrak{K}$ as a split supported by the distance matrix d: we require that for any choice of i, j in $\mathfrak{J}$ and k, l in $\mathfrak{K}$ the sum of the internal distances is at least not the largest among the three distance sums of the quartet i, j, k, l, that is,

$$d_{ij} + d_{kl} < \max\{d_{ik} + d_{jl}, d_{il} + d_{jk}\}; \qquad (1)$$

we then say that $\mathfrak{J}$, $\mathfrak{K}$ is a split with respect to (d_{ij}) or a d-split, for short. Every d-split receives a positive weight, viz., the quantity

$$\alpha_{\mathfrak{J},\mathfrak{K}} = \frac{1}{2} \cdot \min_{\substack{i,j \in \mathfrak{J},\\ k,l \in \mathfrak{K}}} (\max\{d_{ij} + d_{kl}, d_{ik} + d_{jl}, d_{il} + d_{jk}\} - d_{ij} - d_{kl}), \qquad (2)$$

which is called the *isolation index* of $\mathfrak{J}$, $\mathfrak{K}$. All other partitions of the taxa into two parts $\mathfrak{J}$, $\mathfrak{K}$ (that do not qualify as d-splits) thus have index 0. Notice that the isolation index of a split $\mathfrak{J}$, $\mathfrak{K}$ of an ideal tree is exactly the length of the link whose removal results in the two components $\mathfrak{J}$ and $\mathfrak{K}$.

Now, every split $\mathfrak{J}$, $\mathfrak{K}$ gives rise to a *split metric* $\delta_{\mathfrak{J},\mathfrak{K}}$ that assigns distance 1 to two taxa from different parts $\mathfrak{J}$, $\mathfrak{K}$ and zero distance otherwise. As has been proved in Bandelt and Dress (1992), the sum d^1 of all split metrics weighted by their isolation indices with respect to d approximates d from below:

$$d = d^0 + \sum_{\text{splits } \mathfrak{J},\mathfrak{K}} \alpha_{\mathfrak{J},\mathfrak{K}} \cdot \delta_{\mathfrak{J},\mathfrak{K}}, \qquad (3)$$

while the residue $d^0 = d - d^1$ is a metric which does not admit any further splits with positive isolation index. In case of real data the residue d^0 is notoriously nonzero, but still fairly small in comparison to the split-decomposable summand $d^1 = d - d^0$. In order to measure the effectivity of the split decomposition simply compare the average entries of the two matrices d and d^1: the *splittable percentage*

$$\rho := \left(\sum_{\text{taxa } i,j} d^1_{ij} \Big/ \sum_{\text{taxa } i,j} d_{ij} \right) \cdot 100\% \qquad (4)$$

then indicates how much of the given distances between taxa, on the average, is recovered from the weighted sum of split metrics.

To give an example, consider the (artificial) data matrix d for seven taxa A, B, C, D, E, F, G, given in Table 1. The d-splits and isolation indices are readily determined according to the procedure described below. In Table 1 a split such as {A, B, C, D}, {E, F, G} is coded by the shorthand EFG. Since the residue is zero here (that is, $d = d^1$), the given distance between two taxa X and Y equals the sum of the isolation indices corresponding to those code words which contain exactly one of the letters X, Y.

Finding the d-Splits

It is not difficult to compute the d-splits efficiently, since the number of all d-splits is bounded by $\binom{n}{2}$ where n is the number of taxa. One proceeds recursively as follows: enumerate the taxa as $1, 2, \dots, n$, and suppose the d-splits restricted to the subset $\{1, \dots, i - 1\}$ are already determined; then for each d-split $\mathfrak{J}$,

Table 1

A Distance Matrix *d* and Its *d*-Splits (Coded by Their Minority Parts)

	A	B	C	D	E	F
B	4					
C	5	1				
D	7	3	2			
E	13	9	8	6		
F	8	12	13	11	5	
G	6	10	11	13	7	2

Isolation index	Split
6	EFG
4	AFG
2	DEF
1	CDE

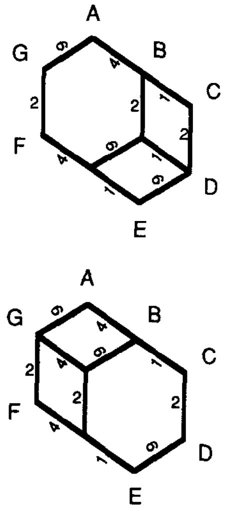

FIG. 1. Two different (equally good) graphical representations of the matrix *d* of Table 1, indicating the four *d*-splits along with their isolation indices.

$\mathfrak{K}$ of this subset check whether $\mathfrak{J} \cup \{i\}$, $\mathfrak{K}$ or $\mathfrak{J}$, $\mathfrak{K} \cup \{i\}$ qualifies as a *d*-split of the enlarged subset $\{1, \ldots, i-1, i\}$; further check whether $\{1, \ldots, i-1\}$, $\{i\}$ is a *d*-split of the enlarged set. This procedure stops after $i = n$ has been processed, providing us with the complete list of *d*-splits of the full set.

The total number of steps is bounded by a polynomial in *n* of degree 6 (with a small leading coefficient). For example, the total number of inequalities (1) that have to be checked in case $n = 8$ is less than 1000 in the worst case and considerably smaller in general, so that this analysis could even be performed by hand for fewer than 10 taxa. A computer program (written in C) is available upon request.

Graphical Representation

The splits of a tree are in one-to-one correspondence with the links and thus are easily read from the diagram. More generally, any split-decomposable metric d^1 can be represented by a mesh-like graph, the links of which are weighted by the corresponding isolation indices. The graphs in question can be chosen among the subgraphs of $\binom{n}{2}$-dimensional cubes (where *n* is the number of taxa), but fortunately they are normally not too weird and can often be drawn in the plane without intersection of links. In contrast to the tree situation, a single split now corresponds to a family of several "parallel" links, which constitutes a cutset, that is, removing these links disconnects ("splits") the graph. Successive application of the following rule determines a cutset: for each "cell," i.e., a cycle without short-cuts, opposite edges belong to the same cutset (and hence receive the same weight). The distance between two taxa *i* and *j* is then obtained as the sum of all weights along a path connecting *i* and *j* which has the smallest number of links. Note that there may be more than one representing graph meeting the requirements and having a minimal number of nodes; see Fig. 1. This graph also appears as a subgraph in Figs. 2 and 4 below.

In order to generate such graphs one proceeds iteratively by incorporating one split after the other: suppose a minimal graph representing a subcollection of splits has been constructed, then this graph is expanded so that the next split gets realized as well, thereby obeying the above rules on cycles, cf. Bandelt (1992). Observe that the order in which the splits are processed may affect the final outcome. For example, the split AFG in the upper graph of Fig. 1 cannot be the last one that gets processed since otherwise the predecessor graph would not have been minimal.

Greedy Tree Selection

If the collection of splits for a matrix *d* is sufficiently large, then it probably includes the splits of trees inferred from the data by other methods. Therefore, in order to estimate a tree, one could select a maximal subset of splits fitting into a tree, so that a certain optimality criterion is met. Indeed, for data sets of medium size the splits obtained from an estimated tree often coincide with the *d*-splits whose indices exceed a certain threshold value.

Recall that a set of splits is realizable on a tree if and only if the splits are pairwise *compatible,* i.e., any two splits $\mathfrak{J}_1$, $\mathfrak{K}_1$ and $\mathfrak{J}_2$, $\mathfrak{K}_2$ of that set have parts, $\mathfrak{J}_1$ and $\mathfrak{J}_2$ say, with empty intersection. For example, every *trivial split,* opposing one taxon to all others, is compatible with all splits.

An optimality criterion would require maximizing an appropriately defined function, e.g., the sum, of the isolation indices (of the chosen splits); optimal solutions could then, of course, be found by branch and bound methods. This bears some resemblance to the compatibility method of Meacham and Estabrook (1985) and the closest tree selection in Hendy and Penny's (1991) spectral analysis.

Even the greedy selection strategy seems to work surprisingly well (when compared to standard methods of tree inference): successively select a new d-split that has the highest isolation index and is still compatible with the splits collected so far. The Sarich (1969) data of immunological distances between eight mammalian species can serve as an illustration. We find exactly 14 d-splits, 13 of which are compatible and thus fit on a tree, while the single "outcast" has minimum isolation index (and is incompatible with the split with maximum index). This is in perfect agreement with the findings of Hendy and Penny (1991) who analyzed the spectrum for these data. The selected tree is, by the way, also in agreement with the one estimated by Fitch (1981), but differs from the one proposed by Yushmanov and Chumakov (1988).

Another good example is provided by the amino acid differences between translated *aroA* gene sequences for nine eubacteria and two eukaryotes; see Table 1 of Griffin and Griffin (1991). We find 19 compatible d-splits plus 3 skew d-splits having much smaller isolation indices (altogether yielding a splittable percentage of more than 92%). This supports the unrooted tree depicted in Fig. 4(b) (Griffin and Griffin, 1991).

In extreme cases, all d-splits turn out to be pairwise compatible and yield the tree obtained by other methods. For example, the archaebacterial tree based on 16S ribosomal RNA sequences, displayed in Fig. 10 of Østergaard *et al.* (1987), is almost perfectly recovered from the d-splits that are indeed pairwise compatible in this case. Only the very short link in the tree, separating the halobacterial pair *Halobacterium volcanii* and *H. morrhuae* from the other archaebacteria (including *H. cutirubrum*), is not recovered by a d-split; this holds for both matrices d given in Fig. 9 (Østergaard *et al.*, 1987), presenting sequence dissimilarities and evolutionary distances, respectively.

RESULTS

Detecting Sequence Convergence

Incompatible d-splits are obtained when split decomposition is applied to distances derived from a set of aligned sequences some of which have undergone massive parallel substitutions. The subsequent cases are instructive: first, parallel amino acid replacements in cow and langur lysozymes, and second, thermophilic convergence in eubacterial ribosomal RNA:

1. Stewart and Wilson (1987) compared the amino acid sequences of lysozymes from cows, langurs, baboons, humans, rats, and horses; cf. Table 4 of Li and Graur (1991, p. 78). They noted that there are four amino acids uniquely shared by cows and langurs. We find six nontrivial splits with respect to amino acid differences: one separating rat, cow, and horse from man and monkeys at an isolation index of 3.0, and five splits each separating a pair of taxa from the rest, viz. rat and horse (with index 7.0), horse and cow (5.5), cow and langur (3.5), langur and baboon (0.5), and baboon and rat (0.5). In particular, the convergence in the cow and langur lineages is manifest in the d-splits, but (with respect to isolation indices) it is less pronounced than the parallelism involving the horse and rat lineages. Since the two nontrivial d-splits with largest indices are incompatible, one concludes that inferring *phylogenetic trees* from these data would not yield reliable results, while concerning parallel evolution they offer rather interesting and valuable information.

2. It has been reported that the GC content in ribosomal RNA sequences of thermophilic bacteria is comparatively high; cf. Li and Graur (1991). This can bias phylogenetic inference, as was clearly demonstrated by Weisburg *et al.* (1989) in the case of 16S ribosomal RNA sequences of 11 eubacteria: De Soete's algorithm, performed for the matrix of evolutionary distances calculated from all gap-free positions, groups together the three thermophiles—*Thermotoga maritima, Thermomicrobium roseum,* and *Thermus aquaticus.* Other distance matrix methods (e.g., ADDTREE and Neighbour-Joining) yield the same result. For a survey on tree reconstruction methods, see Swofford and Olsen (1990).

If only relatively conserved positions are used, then *T. aquaticus* appears to be phyletically closest to *Deinococcus radiodurans;* see Fig. 2 of Weisburg *et al.* (1989).

So we performed split decomposition for both data. The splittable percentage is 73% in either case. The former distances (employing all gap-free positions) yield five nontrivial splits, four of which are compatible. The skew one, the "thermophilic" split separating *T. maritima, T. roseum,* and *T. aquaticus* from the other eubacteria, receives an index of only 2.0, while the *Deinococcus–Thermus* split has an index 11.0. Thus, the greedy tree selection discards the former split. The resulting multifurcation tree is in agreement (after collapsing four links) with tree B from Fig. 2 (Weisburg *et al.*, 1989), estimated from the relatively conserved positions. The latter distances, based on those positions only, give just three compatible splits (occurring in tree B), but lack the "thermophilic" split as well as the *Deinococcus–Thermus* split.

Leffers et al. Data

Leffers *et al.* (1987) compared the 23S ribosomal RNA sequences of six archaebacteria, six eubacteria (including two chloroplasts), and four eukaryotes, and gave the percentages of sequence similarity (based on approximately 2600 nucleotides). These similarity val-

TABLE 2

The Splits along with Their Isolation Indices for the Data Based on Fig. 9 of Leffers *et al.* (1987): Sequence Dissimilarities (per 1000 Positions) and Evolutionary Distances (Estimated Mutational Events per 1000 Positions), Respectively

Isolation index			Isolation index			
Sequence dissimilarity	Evolutionary distance	Split	Sequence dissimilarity	Evolutionary distance	Split	Skew
163.5	214.5	M	100	148.5	CD	
121	144.5	G	99	223.5	MNOP	
116	132	F	84	90	OP	
115	133.5	E	63	75.5	KL	
111	127	N	61.5	123	GHIJKL	
90	102.5	B	60	68	HI	
82	85.5	A	47.5	68	JKL	
65	68	C	46.5	59.5	AB	
64	68.5	D	37.5	47.5	NOP	
56.5	55	J	21.5	31.5	EF	
42.5	44.5	I	11.5	14	ABCDEF	
36.5	33	H	6	2.5	HIJ	+
27	30	K	3.5	1.5	HIJKL	
18.5	18	O	3	0	GHI	+
17.5	12.5	P	2.5	0	JL	+
17	14.5	L	0.5	0	GHIJKLM	+
			0.5	1	CDEF	
			0.5	0	ABOP	+
			0.5	0	DEF	+
			0.5	0	MNO	+
			0.5	0	NO	+

Note. Each split is coded by its minority part. Skew splits (+) are those which are incompatible with at least one split from the estimated phylogenetic tree in Fig. 10 of Leffers *et al.* (1987). Taxon symbols are (A) *Desulfurococcus mobilis;* (B) *Thermoproteus tenax;* (C) *Halococcus morrhuae;* (D) *Halobacterium halobium;* (E) *Methanococcus vanillii;* (F) *Methanobacterium thermoautrophicum;* (G) *Escherichia coli;* (H) *Bacillus stearothermophilus;* (I) *Bacillus subtilis;* (J) *Anacystis nidulans;* (K) *Zea mays* chloroplast; (L) tobacco chloroplast; (M) *Physarum polycephalum;* (N) *Saccharomyces cerevisiae;* (O) *Xenopus laevis;* (P) mouse.

ues were then transformed logarithmically by the Jukes and Cantor method into estimated evolutionary distances. For our reanalysis we used both data sets, where sequence similarities (= percentages of matched positions) were converted into sequence dissimilarities (= percentages of mismatched positions). For convenience, the values of either matrix are multiplied by 10 or 1000, respectively, so that the numbers of mismatched positions and expected mutations, respectively, refer to 1000 positions.

Table 2 below displays the list of splits and isolation indices for both data sets. Figure 2A depicts the graphical composition of all these splits in the case of sequence dissimilarities drawn to scale according to the isolation indices, while Fig. 2B shows the network structure of that graph with all edges given the same length. When passing from sequence dissimilarities to evolutionary distances, the smaller indices of the 16 trivial splits opposing one taxon each to the remaining ones do not change considerably, while the larger ones (say from 90 upward) increase moderately, as is to be expected. There are 21 nontrivial splits for the sequence dissimilarities, 10 of which receive indices smaller than 10. From the latter ones 7 drop out (i.e., get index 0) when evolutionary distances are considered instead; on the other hand, two indices increase drastically, viz., for the splits separating the eukaryotes or the eubacteria, respectively, from the other taxa. The split separating the archaebacteria from the two other kingdoms has a fairly low isolation index: 11.5 and 14, respectively. This is not surprising in view of the ongoing controversy about the phylogenetic status of the archaebacteria; see Kjems and Garrett (1990) for a recent contribution. On the other hand, no split from Table 2 would reject the monophyletic status of archaebacteria. Observe that for either data matrix all splits are present (with positive isolation index) that correspond to the links of the estimated phylogenetic tree shown in Fig. 10 of Leffers *et al.* (1987); the other splits in our Table 2 are marked as "skew" (to this reference tree).

This tree is returned by the greedy tree selection from the splits with respect to either matrix. In the case of evolutionary distances, there is only one skew split, viz. HIJ (that is, the split which separates {H, I, J} from the rest), which is incompatible with the split JKL having index 68. That *Anacystis nidulans* (J) and the two *Bacilli* (H, I) give rise to a split may partially

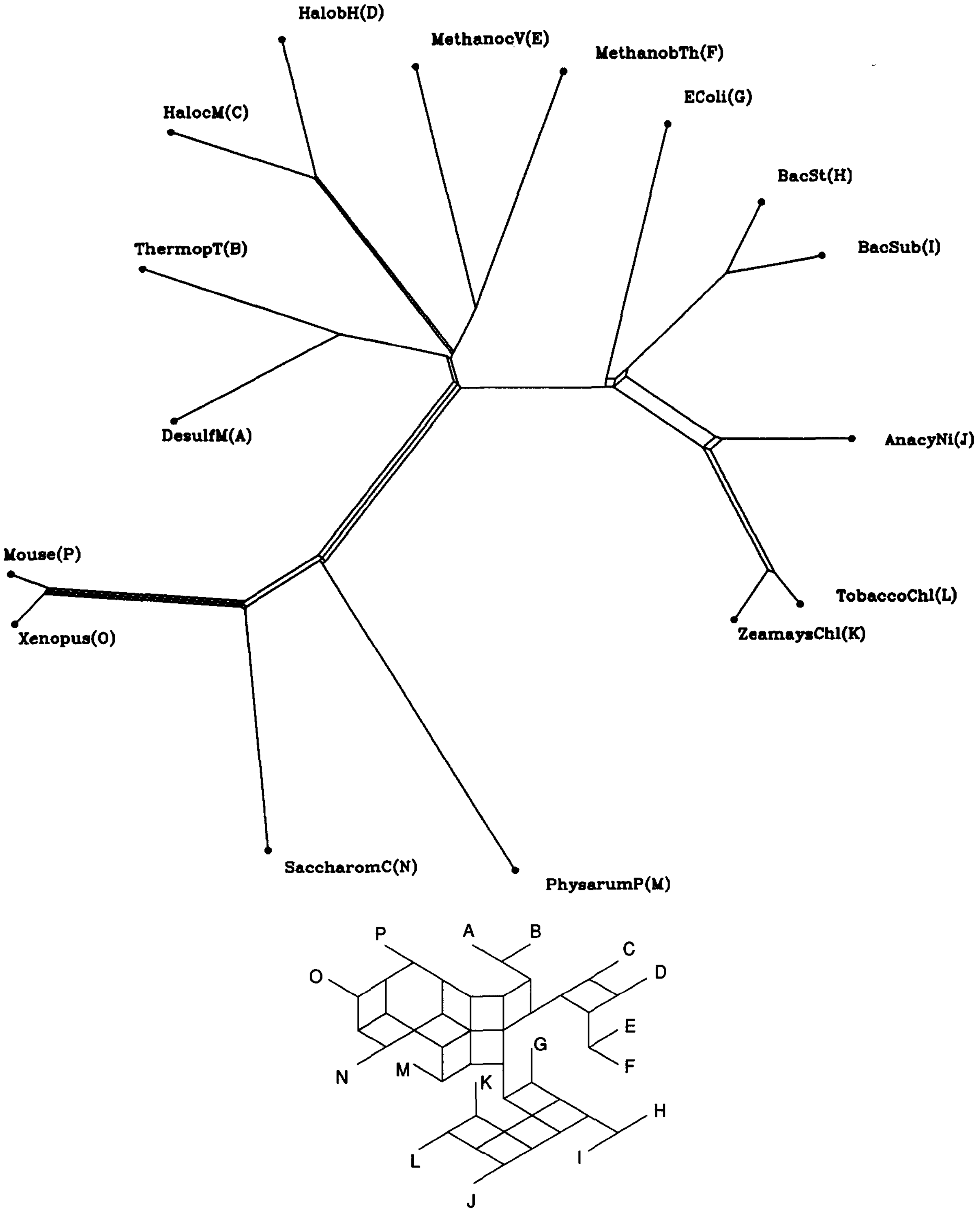

FIG. 2. A. A graphical representation of the split-decomposable part of the matrix of sequence dissimilarities according to Table 2 with edge lengths proportional to the isolation indices (drawn to scale). B. A graph, isomorphic to the graph in A, but with all edges given the same length. Taxon symbols are as in A and Table 2.

be explained by the fact that the other three members of the eubacterial kingdom, *Escherichia coli* (G) and the two chloroplasts (K, L), show a higher degree of mutability. Summarizing, we can state that according to the isolation indices of splits the evolutionary distances clearly support a unique tree, the only skew split being HIJ with index 2.5. A similar conclusion could also be drawn from the reanalysis of the same data set in Bandelt and Dress (1989), where "weak clusters" were employed.

When one focusses on a particular subgroup of taxa, e.g., the archaebacteria in the study of Leffers *et al.* (1987), then the corresponding distance submatrix should be investigated separately. Since a single quartet of taxa, two of which are in either part of a potential split, can cause the rejection of this split (see the definition of isolation index), the total number of splits (with positive index) tends to be relatively small for larger data sets. So, some of the "local" information on parallelism and systematic error reflected in the distance matrix for a subgroup of taxa is lost in split analysis (i.e., transferred to the residue) when other, distantly related taxa are taken into account.

We therefore computed the splits and their indices for those 6 archaebacteria: for either distance type, 14 splits are obtained; 5 of the 8 nontrivial splits are skewed with respect to the reference tree. The residue here is easy to describe: it is associated with the graph $K_{2,3}$, in which two nodes are linked to three other nodes respectively such that all six links are of the same length. It is thus not difficult to present a geometric picture from which the splits as well as the residue are conveniently read off, thereby recovering the distances as the lengths of shortest paths; see Fig. 3. The skew split with largest index (viz., 5.5 or 10, respectively) is ABF (alias CDE), which separates the thermophiles *Desulfurococcus mobilis* (A), *Thermoproteus tenax* (B), and *Methanobacterium thermoautrophicum* (F) from the other three archaebacteria.

The same type of analysis can be performed for, say, the six eubacteria. As skew splits one then obtains HIJ, IJ, and JL with respect to either distance measure. Observe that the residue is much larger for the evolutionary distances. A similar observation can be made for several other subsets of taxa; see Table 3. The transformation from sequence dissimilarities to evolutionary distances increases the residual parts, while the number of skew splits decreases. There is thus a trade-off between indecomposable "noise" and the incompatibility between splits for these data, which needs further analysis.

Weisburg et al. Data

Remarkably, it is not always true that the number of splits for estimated evolutionary distances is smaller than the one for uncorrected sequence dissimilarities. The 16S ribosomal RNA data for 10 species of eubacteria, provided by Weisburg *et al.* (1991), constitute an exception: the evolutionary distances admit 25 splits (shown in our Fig. 4) including all 17 splits from the estimated tree (in Fig. 3, Weisburg *et al.*, 1991), while two of the skew splits (viz. ACD and ADE) are not present in the dissimilarity data. The splittable percentages are 88.9% (evolutionary distance) and 83.3% (dissimilarity), respectively. It may be noted that the Jukes and Cantor transformation increases also the variation in these data. This correlates with the high isolation degrees of two incompatible splits, viz. ADEFG with index 13.5 (per 1000 sites) and AC with index 9, while the corresponding indices equal 2 and 1 in the case of uncorrected dissimilarities.

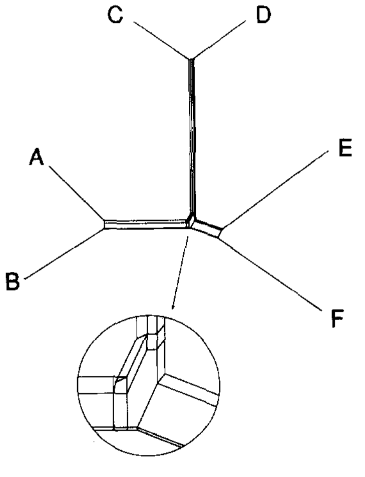

FIG. 3. Exact representation of the evolutionary distances between the six archaebacteria (drawn to scale).

The greedy tree selection recovers the tree proposed by Weisburg *et al.* (1991) when evolutionary distances are used. For the sequence dissimilarities another tree results, where the pair *Rhodospirillum rubrum* and *Rhodopseudomonas palustris* would branch off first (given *E. coli* as an outgroup). In any case, the deeper branchings in both trees seem to be somewhat uncertain.

Figure 5 demonstrates how the computer graphics program based on our theory and written by Rainer Wetzel step by step creates the network shown in Fig. 4B (taxa symbols are as in Fig. 4A) by incorporating one split after another into the evolving network. The following splits have been incorporated consecutively: ABCI–DEFGHJ, ADE–BCIFGHJ, ACD–BEFGHIJ, ACDE–BFGHIJ, ABC–DEFGHIJ, ADEFG–BCFGHIJ, ABCHIJ–DEFG, ACDEFG–BHIJ, ADEFGHIJ–BC, AC–BDEFGHIJ, AD–BCEFGHIJ, ABCFGHIJ–DE, ABCDEFG–HIJ, ABCDEFGJ–HI, ABCDEIJH–FG, and—in the last network—all splits one versus rest.

Huss and Sogin Data

In contrast to the latter data set, the system of splits for the 18S ribosomal RNA data investigated by Huss

TABLE 3

The Relative Sizes of the Split-Decomposable Parts and the Numbers of All Splits and All Skew Splits, Respectively, for Various Subsets of Taxa

Splittable percentage of the total distance		No. all splits : No. skew splits			
Sequence dissimilarity	Evolutionary distance	Sequence dissimilarity	Evolutionary distance	No. taxa	Taxon groups
87.9	82.2	37:8	30:1	16	All three kingdoms
91.2	88.6	28:7	27:6	12	Archaebacteria and eubacteria
95.3	92.1	25:8	21:4	10	Archaebacteria and eukaryotes
95.9	94.1	26:9	22:5	10	Eubacteria and eukaryotes
98.0	97.0	12:3	12:3	6	Eubacteria
99.3	98.8	14:5	14:5	6	Archaebacteria
100	100	6:1	6:1	4	Eukaryotes

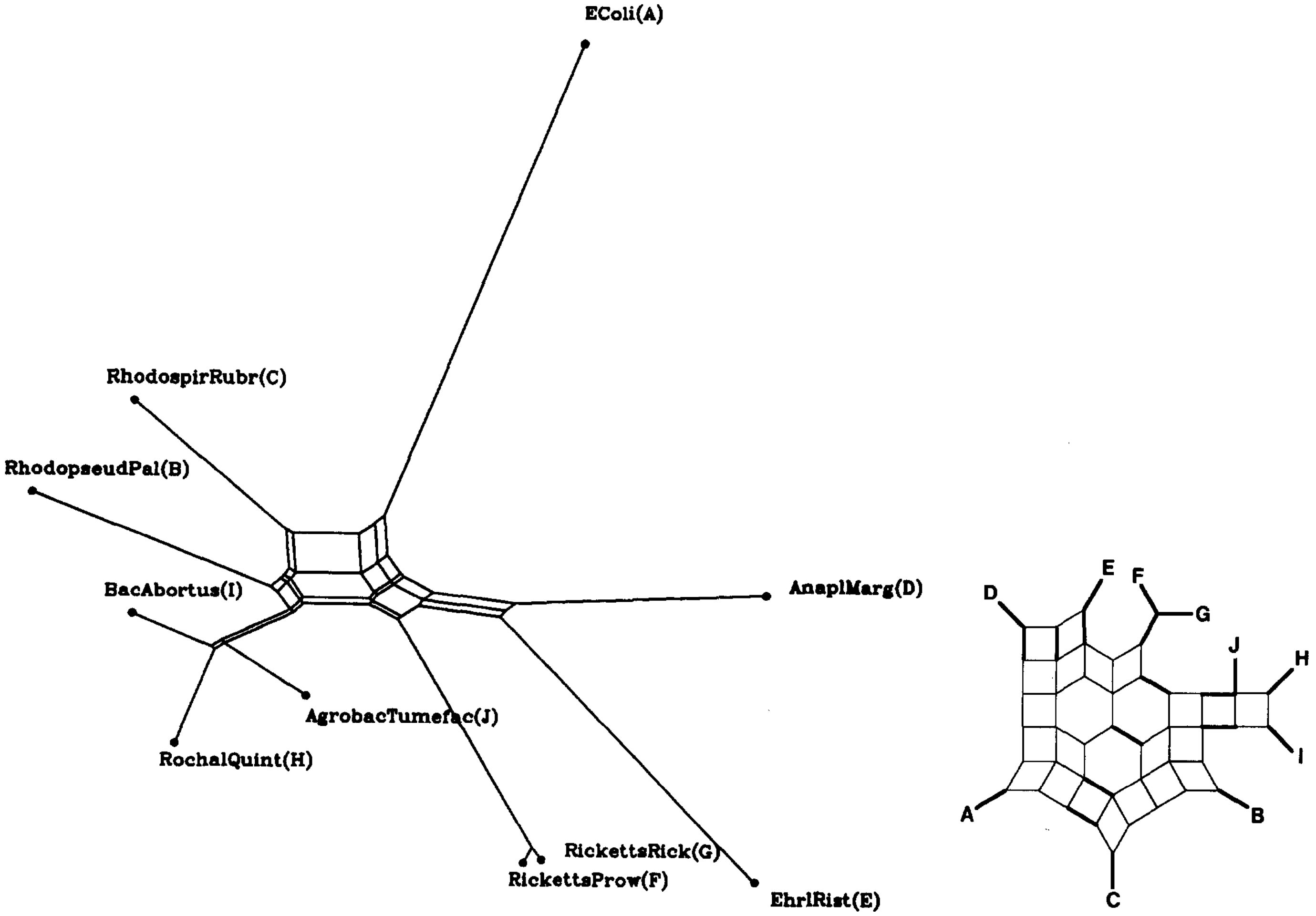

FIG. 4. A. A graphical representation of the split-decomposable part of the evolutionary distances between Rickettsiales and other eubacteria drawn to scale (data from Table 4 of Weisburg *et al.* (1991)). Taxon symbols are: (A) *Escherichia coli;* (B) *Rhodopseudomonas palustris;* (C) *Rhodospirillum rubrum;* (D) *Anaplasma marginale;* (E) *Ehrlichia risticii;* (F) *Rickettsia prowazekii;* (G) *Rickettsia ricketsii;* (H) *Rochalimaea quintana;* (I) *Bacillus abortus;* (J) *Agrobacterium tumefaciens.* B. A graph, isomorphic to the graph in A, but with all edges given the same length. Bold lines indicate links corresponding to the splits with isolation index larger than 10 (per 1000 sites). Taxon symbols are as in A.

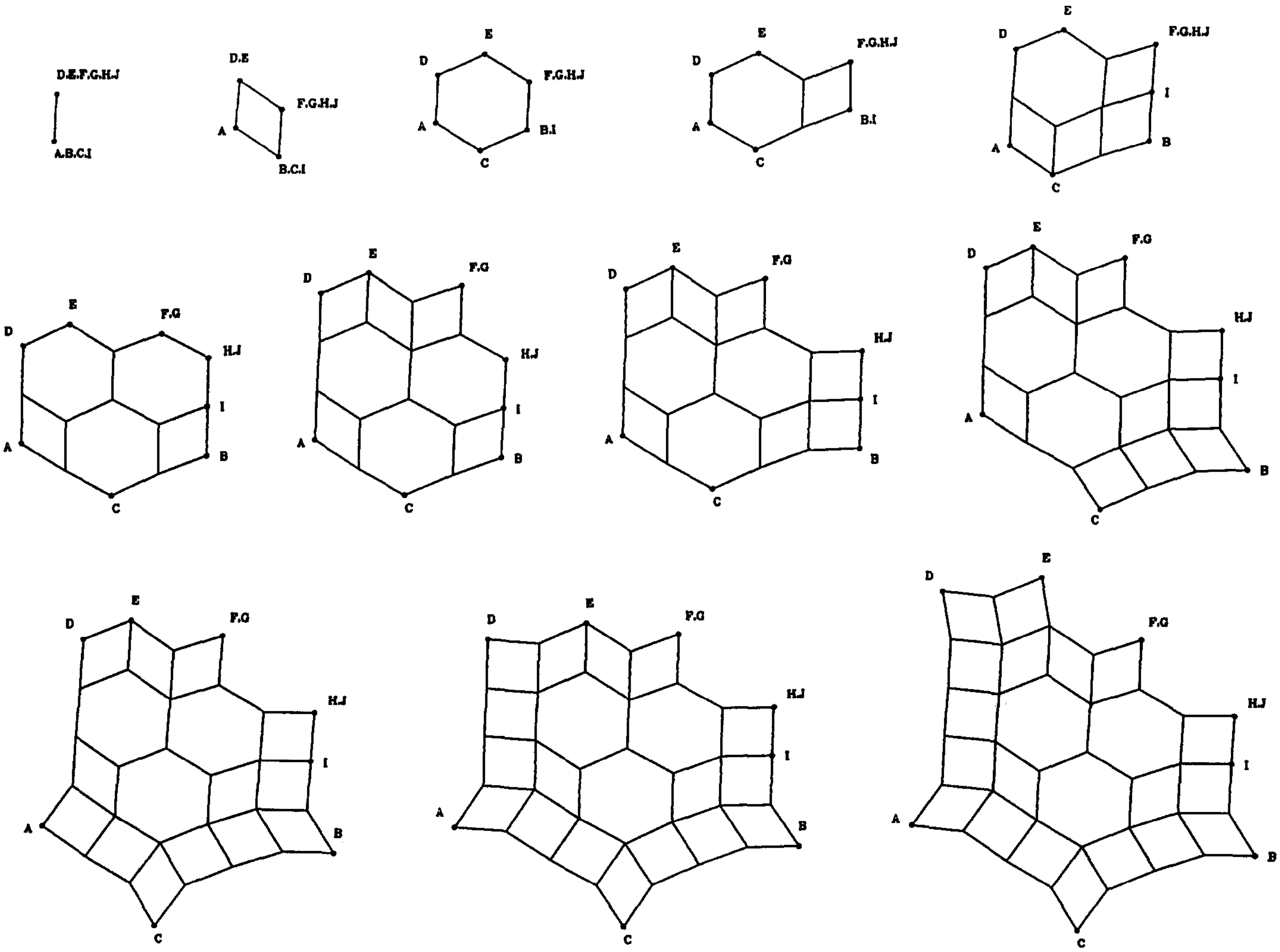

FIGURE 5

and Sogin (1990) is easier to interprete—at the expense of having smaller splittable percentages, 79.9% (dissimilarity) and 78.8% (evolutionary distance). Isolation indices do not change considerably here when passing from sequence dissimilarities to evolutionary distances (estimated by the Jukes and Cantor method); see Table 4. The greedy tree selection, departing from the splits with respect to sequence dissimilarities, yields the tree displayed in Fig. 2 of Huss and Sogin (1990), except that the link separating IJKL from the rest (the shortest one on that tree) is not recovered by a *d*-split; it is not even supported by a partial *d*-split as shown by a secondary analysis performed only for the Chlorococcales (including *Nanochlorum*). In the case of evolutionary distances the tree split ABCDEF (separating higher plants and *Chlamydomonas* from the other green algae) is lost. One of the skew splits, viz. ABCDEH with index 1.5 (per 1000 sites) for either type of distance, can be interpreted as a distance artifact: it splits higher plants (the outgroup organisms) together with the "fast-clock" organism *Chlorella protothecoides* from the other green algae. This split is unlikely to bear phylogenetic information as it is incompatible with two (compatible) splits receiving indices 8.5 and 4.5 (dissimilarity)/4 (evolutionary distance), respectively.

CONCLUSIONS

Split decomposition can enhance phylogenetic analysis of distance data by detecting opposite groupings ("splits") of organisms that are defined by distinctive distance features, caused by common ancestry, convergence, or systematic or random errors. A major part of random noise contained within the data is transferred to the split-prime residue, which is removed from the data in the course of analysis. This residue typically covers 10 to 30% of the total distance in the case of ribosomal RNA data (with about 10 to 25 taxa), whereas for randomly chosen metrics this amount eas-

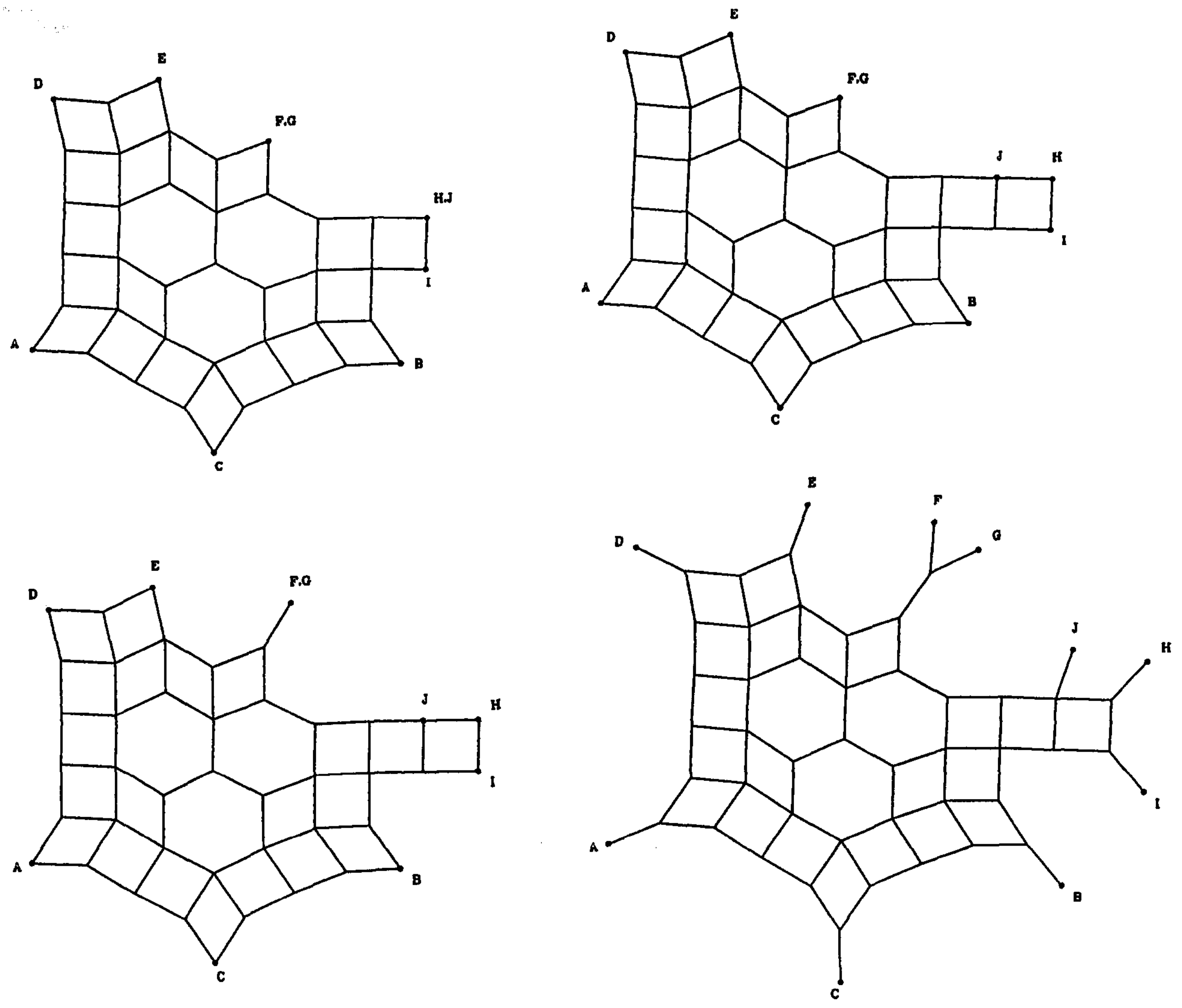

FIG. 5—*Continued*

ily exceeds 50%. Some portion of random and systematic error survives in the split-decomposable part and is manifest in the incompatibilities of splits. A split is likely to fall into this category when its isolation index is relatively small and it is incompatible with splits having much larger indices. Therefore, selecting a clique of compatible splits in a greedy fashion according to isolation indices often recovers most of the phylogenetic trees that are estimated by other methods, but normally leaves unresolved the most uncertain furcations.

A graphical representation of the split-decomposable part of a distance matrix furthers the understanding of tentative phylogenetic relationships plus inherent parallelism. If the number of taxa is very small, then it is possible sometimes even to integrate the split-prime residue into the diagram as well, thus visualizing the full decomposition of the distance matrix.

For very large data sets that include fairly distant groups of taxa, one can additionally perform a secondary analysis of "partial *d*-splits." To this end the whole set of taxa is partitioned into smaller subcollections identified by compatible splits with large isolation indices.

When several distance matrices for one and the same set of organisms are available (for instance, through different weighting schemes of characters or particular methods of correction), it can be instructive to compare the corresponding splittable percentages and the structure of the resulting split systems, in order to evaluate the phylogenetic content of the respective distance matrices.

In closing, we may speculate on further potential applications of our method. In view of its ability to process incompatible splits, split decomposition perhaps can also serve as a tool for investigating reticulate evolution. It is, however, not obvious how to clearly discriminate between random and systematic

TABLE 4

The Nontrivial Splits along with Their Isolation Indices (per 1000 Sites) for the Sequence Dissimilarities and Evolutionary Distances, Respectively, between Green Algae and Higher Plants (Data from Huss and Sogin (1990))

Isolation index			
Sequence dissimilarity	Evolutionary distance	Split	Skew
41.5	48	ABCDE	
14	14	NO	
13.5	14	BC	
9.5	10.5	BCDE	
8	8.5	GH	
4.5	4.5	MNO	
4.5	4.0	GHIJKL	
4	3.5	DE	
1.5	1.5	ABCDEH	+
1	1	AD	+
1	0.5	MO	+
1	0.5	KL	
0.5	1	IJ	
0.5	0	ABCDEF	

Note. The nonskew splits fit in the estimated phylogenetic tree in Fig. 2 (ibidem). Each split is coded by its minority part. Taxon symbols are (A) *Zamia pumila;* (B) *Oryza sativa;* (C) *Zea mays;* (D) *Lycopersicon esculentum;* (E) *Glycine max;* (F) *Chlamydomonas reinhardtii;* (G) *Prototheca wickerhamii;* (H) *Chlorella protothecoides;* (I) *Chlorella minutissima;* (J) *Nanochlorum eucaryotum;* (K) *Chlorella vulgaris;* (L) *Chlorella kessleri;* (M) *Ankistrodesmus stipitatus;* (N) *Chlorella fusca;* (O) *Snedesmus obliquus.*

error on the one hand and hybridization events on the other. It is equally difficult, at the population level, to discriminate between dendritic evolution and evolutionary process conforming to the quasispecies model (compare, for example, the fine meshed diagram relating some eubacterial species shown in Fig. 4 above with the less complex Fig. 4b from Dopazo *et al.* (1990), depicting mutual relationships between certain viruses).

ACKNOWLEDGMENTS

We thank William Martin and Rainer Wetzel for their help in generating some of the figures and for comments on an earlier version of the manuscript.

REFERENCES

Bandelt, H.-J. (1992). Generating median graphs from Boolean matrices. *In* L_1-Statistical Analysis and Related Methods (Y. Dodge, Ed.), pp. 305–309, Elsevier, Amsterdam.

Bandelt, H.-J., and Dress, A. W. M. (1989). Weak hierarchies associated with similarity measures—An additive clustering technique. *Bull. Math. Biol.* **51:** 133–166.

Bandelt, H.-J., and Dress, A. W. M. (1992). A canonical decomposition theory for metrics on a finite set. *Advances Math.* **92:** 47–105.

Dopazo, J., Dress, A., and von Haeseler, A. (1990). Split decomposition: A new technique to analyse viral evolution. Preprint 90-037, SFB 343, Universität Bielefeld.

Fitch, W. M. (1981). A non-sequential method for constructing trees and hierarchical classifications. *J. Mol. Evol.* **18:** 30–37.

Griffin, H. G., and Griffin, A. M. (1991). Cloning and DNA sequence analysis of the *serC–aroA* operon from *Salmonella gallinarum;* evolutionary relationships between the prokaryotic and eukaryotic *aroA*-encoded enzymes. *J. Gen. Microbiol.* **137:** 113–121.

Hendy, M. D. (1989). The relationship between simple evolutionary tree models and observable sequence data. *Systemat. Zool.* **38:** 310–321.

Hendy, M. D., and Penny, D. (1991). Spectral analysis of phylogenetic data. Preprint.

Huss, V. A. R., and Sogin, M. L. (1990). Phylogenetic position of some *Chlorella* species within the Chlorococcales based upon complete small-subunit ribosomal RNA sequences. *J. Mol. Evol.* **31:** 432–442.

Kjems, J., and Garrett, R. A. (1990). Secondary structural elements exclusive to the sequences flanking ribosomal RNAs lend support to the monophyletic nature of the archaebacteria. *J. Mol. Evol.* **31:** 25–32.

Leffers, H., Kjems, J., Østergaard, L., Larsen, N., and Garrett, R. A. (1987). Evolutionary relationships amongst archaebacteria. A comparative study of 23S ribosomal RNAs of a sulphur-dependent extreme thermophile, an extreme halophile, and a thermophilic methanogen. *J. Mol. Biol.* **195:** 43–61.

Li, W.-H., and Graur, D. (1991). Fundamentals of molecular evolution. Sinauer Associates, Sunderland, MA.

Meacham, C. A., and Estabrook, G. F. (1985). Compatibility methods in systematics. *Anno. Rev. Ecol. Syst.* **16:** 431–446.

Østergaard, L., Larsen, N., Leffers, H., Kjems, J., and Garrett, R. (1987). A ribosomal RNA operon and its flanking region from the archaebacterium *Methanobacterium thermoautotrophicum,* Marburg strain: Transcription signals, RNA structure and evolutionary implications. *Syst. Appl. Microbiol.* **9:** 199–209.

Sarich, V. M. (1969). Pinniped origins and the rate of evolution of carnivore albumins. *Systemat. Zool.* **18:** 286.

Stewart, C.-B., and Wilson, A. C. (1987). Sequence convergence and functional adaption of stomach lysozymes from foregut fermenters. *Cold Spring Harbor Symp. Quant. Biol.* **52:** 891–899.

Swofford, D., and Olsen, G. (1990). Phylogeny reconstruction. *In* Molecular Systematics (D. M. Hillis and C. Moritz, Eds.), pp. 411–501, Sinauer Associates, Sunderland, MA.

Weisburg, W. G., Barns, S. M., Pelletier, D. A., and Lane, D. J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.* **173:** 697–703.

Weisburg, W. G., Giovannoni, S. J., and Woese, C. R. (1989). The *Deinococcus–Thermus* phylum and the effect of rRNA composition on phylogenetic tree construction. *Syst. Appl. Microbiol.* **11:** 128–134.

Yushmanov, S. V., and Chumakov, K. M. (1988). Algorithms for constructing phylogenetic trees of maximum similarity. *Mol. Genet. Mikrobiol. Virusol.* **3:** 9–15 [in Russian].