Inference of Horizontal Genetic Transfer From Molecular Data: An Approach Using the Bootstrap

Jeffrey G. Lawrence[1] and Daniel L. Hartl

Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110

Manuscript received June 28, 1991
Accepted for publication March 31, 1992

ABSTRACT

Inconsistencies in taxonomic relationships implicit in different sets of nucleic acid sequences potentially result from horizontal transfer of genetic material between genomes. A nonparametric method is proposed to determine whether such inconsistencies are statistically significant. A similarity coefficient is calculated from ranked pairwise identities and evaluated against a distribution of similarity coefficients generated from resampled data. Subsequent analyses of partial data sets, obtained by the elimination of individual taxa, identify particular taxa to which the significance may be attributed, and can sometimes help in distinguishing horizontal genetic transfer from inconsistencies due to convergent evolution or variation in evolutionary rate. The method was successfully applied to data sets that were not found to be significantly different with existing methods that use comparisons of phylogenetic trees. The new statistical framework is also applicable to the inference of horizontal transfer from restriction fragment length polymorphism distributions and protein sequences.

THE accumulation of nucleic acid sequences from diverse taxa has prompted phylogenetic inference based upon a wide range of molecular relationships (FIELD *et al.* 1988; LAWRENCE, OCHMAN and HARTL 1991; THOMAS *et al.* 1989; WOESE 1987). Inconsistencies are sometimes found in the taxonomic relationships inferred from the nucleotide sequences of different genes. These differences may result from variation in evolutionary rates within and among genes, high levels of homoplasy due to convergent evolution, or events in which genetic material has been transferred between distantly related taxa. Horizontal transfer of genetic material represents a powerful mechanism by which organisms may acquire novel metabolic pathways or substantially alter existing pathways. Many cases of possible gene transfers between distantly related organisms or organelle genomes have been suggested, typically motivated by the finding of unusual similarities among homologous sequences from distantly related taxa, or by observing inconsistencies in gene phylogenies from a group of of organisms (*e.g.*, BANNISTER and PARKER 1985; BRISSON-NOEL, ARTHUR and COURVALIN 1988; CARLSON and CHELM 1986; DOOLITTLE *et al.* 1990; DOWSON *et al.* 1989; HILDEBRANDT *et al.*. 1989; IWAASA, TAKAGI and SHIKAMA 1989; LANDAN *et al.* 1990; LIAUD, ZHANG and CERFF 1990; PEÑALVA *et al.* 1990; PLOS *et al.* 1989; WAKABAYASHI, MATSUBARA and WEBSTER 1986; WRIGHT and CUMMINGS 1983). Since it has been difficult to provide convincing statistical support for horizontal gene transfer, several cases have been disputed, often by the same data being interpreted in an alternate manner (*e.g.*, KOLL 1986; LEUNISSEN and DE JONG 1986; SHATTERS and KAHN 1989). Examination of the evolutionary histories of organelle genomes has also suggested possible genetic transfer between organelle genomes and the nucleus (BALDAUF and PALMER 1990; ELLIS 1982; FARELLY and BUTOW 1983; FUKADA *et al.* 1985; GELISSEN *et al.* 1983; LIAUD, ZHANG and CERFF 1990; MARÉCHAL-DROUARD *et al.* 1990).

To attribute inconsistencies in molecular data to horizontal genetic transfer, one must first determine whether the taxonomic relationships implicit in two data sets are significantly different. Once a significant difference has been established, one must distinguish between differences due to homoplasy or variation in evolutionary rates from differences resulting from horizontal transfer of genetic material. The problem, as DOW and CHEVERUD (1985) noted, is that few formal procedures have been developed to assess the statistical significance of differences in sets of relationships. The test may be accomplished in an indirect fashion by comparing phylogenies inferred from the data. However, existing methods for testing phylogenies were designed to ascertain which of many possible phylogenies is best supported by a single data set (*e.g.*, FELSENSTEIN 1981, 1985, 1988; NEI, STEPHENS and SAITOU 1985; PENNY and HENDY 1986; SNEATH 1986; TEMPLETON 1983a,b, 1985). These tests have also been extended to compare the phylogenies inferred from different data sets. However, tests utilizing phylogenies are intimately tied to particular genealogies, and they may be compromised by the inability of particular trees, or tree-making algorithms,

[1] Current address: Department of Biology, University of Utah, Salt Lake City, Utah 84112.

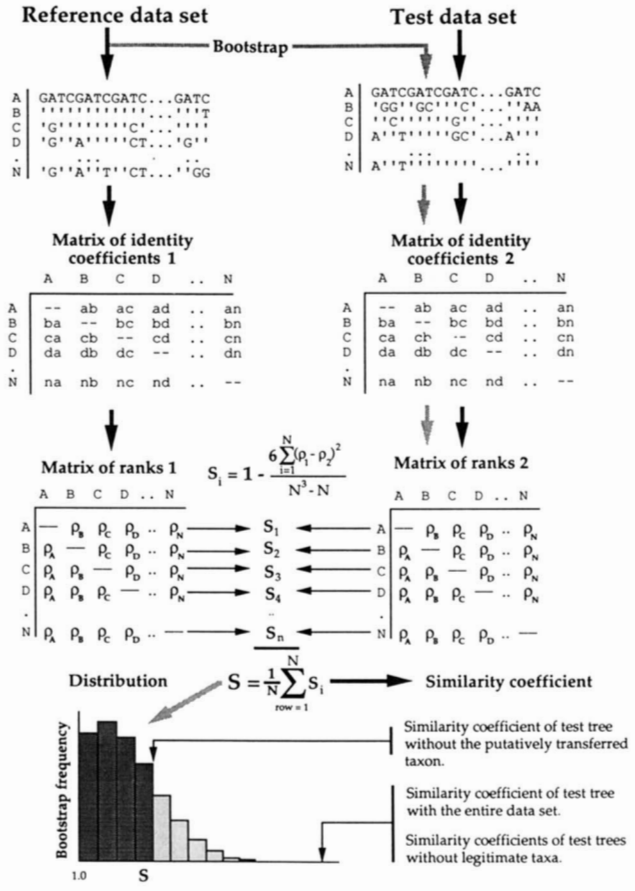

FIGURE 1.—Flow chart describing the similarity coefficient and the generation of the test distribution by resampling.

to embrace the full range of possible relationships implicit in the data.

We have developed a method that tests relationships implicit in nucleotide sequences without the *a priori* construction of phylogenetic trees. This approach addresses the fundamental issue of whether relationships implicit in two sets of nucleotide sequences are significantly different, independent of issues concerning which, if any, inferred phylogeny best describes each data set. Furthermore, when two sets of nucleotide sequences are found to be significantly different, subsequent analyses of subsets of sequences can often identify the taxon to which the difference may be attributed. The phylogeny that best describes each data set then remains a separate issue.

METHODS

Statistical test: Figure 1 illustrates the method for testing potential differences in taxonomic relationships implicit in two sets of nucleotide sequences. First, a matrix of identity coefficients is generated for each data set, in which the entry in cell (i, j) is the proportion (or percentage) of sites that are identical between taxa i and j (calculated as the number of identical sites, divided by the average number of sites, between i and j). Each row in the matrix of identity coefficients represents a set of comparisons between a particular taxon and all the other taxa. To make the matrices commensurate, the magnitudes of the relationships are ranked within each row to create corresponding matrices of ranks; ties are assigned midrank values. The two matrices of ranks are compared row by row using the Spearman rank correlation statistic (SIEGEL 1956). Spearman statistics range from +1.0, indicating a perfect positive correlation between the ranks, to −1.0, indicating a perfect negative correlation between the ranks. The Spearman statistics for all rows in the matrices of ranks are averaged to yield an overall similarity coefficient. Similarity coefficients of +1.0 indicate identical taxonomic relationships implicit in the two sets of data, while those less than +1.0 denote some discrepancy in the relationships.

To assess the statistical significance of a discrepancy, we employ the bootstrap, a method of resampling data designed to simulate the variability of a particular estimate (EFRON 1979). This approach has been utilized to assign confidence limits to particular nodes of phylogenetic trees (FELSENSTEIN 1985). In the bootstrap, individual observations from the original data set (in our case aligned positions in a set of nucleotide sequences) are chosen at random with replacement to create a simulated data set with the same sample size. The parameter of interest is estimated from the simulated data, and the process is repeated numerous times to approximate the sampling distribution of the estimate. In our application, nucleotide positions are sampled at random to generate a simulated data set. Matrices of identity coefficients and matrices of ranks are generated, and a similarity coefficient between the bootstrapped matrix of ranks and the original matrix of ranks is computed. This process is repeated to compile a distribution of similarity coefficients. This distribution is examined to determine if the magnitude of the similarity coefficient of the two original data sets is statistically significant, that is, if it lies outside the distribution of similarity coefficients generated from the bootstrapped data sets. The proportion of bootstrapped simulations with smaller similarity coefficients than that of the two original data sets is taken as the significance value in a one-tailed test. In a sense, this application of the bootstrap addresses the question: "Are the genealogical relationships implicit in one data set consistent with those implicit in another data set?"

For cases of horizontal genetic transfer, the similarity coefficient between two sets of nucleotide sequences should be significantly smaller than the distribution of similarity coefficients generated from the bootstrapped data sets. However, in the case of horizontal transfer, excluding a particular taxon should eliminate the discrepancy, while excluding any other taxon would not. This procedure requires comparing significance levels, and it should therefore be consid-

ered as a consistency check when there is an *a priori* hypothesis of horizontal transfer rather than as a method for detecting horizontal transfer. In contrast to the situation with horizontal transfer, data sets exhibiting high levels of homoplasy may be found to be significantly different, but the elimination of one taxon, or any combination of taxa, would not be expected to eliminate the discrepancy. Moderate levels of homoplasy may be accounted for in our method by relaxing the criteria by which ties are assigned for ranking purposes (see below). In some cases it might make a difference which data set is bootstrapped, hence significance should be corroborated by retesting after interchange of reference and test data sets. In this manner, levels of homoplasy or variation in evolutionary rate that differ between the data sets will not bias the significance of the test.

It should be emphasized here that interchange of reference and test data sets, or retesting after eliminating one or more taxa from the data, always requires adjustment of the significance level in order to compensate for the performance of multiple tests.

In principle, using the ranks of the values in each row to compare the matrices of identity coefficients loses statistical power relative to using the identity coefficients themselves. In order to reflect the actual identity coefficients more accurately, it may be advantageous in some cases to relax the criteria by which ties in rank are assigned. For example, if the identity coefficients of taxon A relative to B, C, D, E and F are 91, 90, 70, 32 and 30, respectively, the ranks are 1, 2, 3, 4 and 5. However, because of sampling variation, additional data from the same taxa may lead to identity coefficients of 89, 90, 74, 31 and 33, and in this case the ranks would be 2, 1, 3, 5 and 4. Clearly, in a case like this, it would be preferable to assign A-B and A-C as approximately equal in rank, and similarly for A-E and A-F, hence the ranking would be 1.5, 1.5, 3, 4.5 and 4.5, respectively.

Data sets: Simulated data sets were derived using the aligned nucleotide sequences of the *gap* locus, encoding glyceraldehyde-3-phosphate dehydrogenase, from enteric bacteria (LAWRENCE, OCHMAN and HARTL 1991). Alternate 10-nucleotide sections were partitioned into two data sets. In the reference data sets, the taxonomic relationships are those determined from the nucleotide sequences of *gap*, encoding the glycolytic enzyme glyceraldehyde-3-phosphate dehydrogenase, and of *ompA*, encoding an outer membrane protein (LAWRENCE, OCHMAN and HARTL 1991). Simulated horizontal transfer ("transfer data sets") were derived from the reference data sets by substituting the *Escherichia fergusonii* (Efe) sequence for the *Serratia odorifera* (Sod) sequence. Data for the Drosophila transposable element *mariner* and for the *Adh* locus, encoding alcohol dehydrogenase, were taken from MARUYAMA and HARTL (1991) and JEFFS and ASHBURNER (1991).

TABLE 1

Pairwise divergences at nonsynonymous sites within reference data sets (upper diagonal) and transfer data sets (lower diagonal)

	Eco	Sty	Evu	Kpn	Sod
Eco	—	1.2	3.6	3.3	8.1
Sty	1.2	—	3.2	3.0	7.5
Evu	3.6	3.2	—	3.0	6.9
Kpn	3.3	3.0	3.0	—	7.1
Sod	0.8	0.9	3.1	2.7	—

Data from LAWRENCE, OCHMAN and HARTL (1991); pairwise divergences were corrected for multiple substitution by the method of PERLER *et al.* (1980). Taxon designations are: Eco, *Escherichia coli*; Efe, *Escherichia fergusonii*; Sty, *Salmonella typhimurium*; Evu, *Escherichia vulneris* ATCC 33821; Kpn, *Klebsiella pneumonia*; Sod, *Serratia odorifera*. Taxon Sod in the transfer set is assigned nucleotide sequences derived from Taxon Efe.

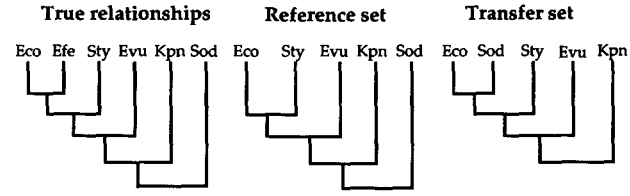

FIGURE 2.—Phylogenetic representation of the taxonomic relationships inferred from the reference and transfer data sets of the simulated horizontal transfer, as compared to the true relationships (after LAWRENCE, OCHMAN and HARTL 1991). Taxon designations as in Table 1.

RESULTS

As an initial test of the method, we created simulated data sets and introduced aberrations in the nucleotide sequences that mimicked a horizontal transfer event. Four sets of nucleotide sequences were derived from sequences of *gap* genes from enteric bacteria. The taxonomic relationships implicit in the reference and transfer data sets are summarized in Table 1. These relationships may be represented by the phylogenies shown in Figure 2. (It should be emphasized that trees are only convenient for representing groups of relationships. The test is independent of trees, and indeed, the matrices of ranks may be significantly different even though tree-making algorithms fail to deduce a unique topology.) While the nucleotide sequences of the reference data sets define one set of taxonomic relationships [consistent with the *gap* and *ompA* loci (LAWRENCE, OCHMAN and HARTL 1991)], the nucleotide sequences of the transfer data sets define a substantially different set of relationships. This discrepancy was introduced by assigning the nucleotide sequence of taxon Efe to taxon Sod in the transfer data set. Each reference data set was then compared to the transfer data set derived from the alternate portions of the *gap* genes, and similarity coefficients were computed and tested. The two ref-

TABLE 2

Data for simulated horizontal transfer event, 1000 bootstrap repetitions

Bootstrap data set[a]	Comparison data set	Taxon eliminated[b]	Distribution		Similarity coefficient	P^c
			μ	σ		
Ref1	Ref2		0.948	0.029	0.96	0.773
Ref2	Ref1		0.900	0.100	0.96	0.900
Tra1	Tra2		0.889	0.061	0.96	0.948
Tra2	Tra1		0.863	0.066	0.96	0.966
Ref1	Tra2		0.946	0.031	0.56	0.000
		Eco	0.944	0.044	0.65	0.000
		Sty	0.957	0.039	0.65	0.000
		Evu	0.968	0.033	0.60	0.000
		Kpn	0.964	0.040	0.40	0.000
		Sod	0.913	0.064	0.95	0.855
Ref2	Tra1		0.898	0.061	0.56	0.000
		Eco	0.921	0.065	0.60	0.000
		Sty	0.900	0.072	0.65	0.000
		Evu	0.930	0.067	0.60	0.000
		Kpn	0.918	0.070	0.45	0.000
		Sod	0.882	0.085	0.95	0.880
Ref1	Omp		0.968	0.034	1.00	1.000
Ref2	Omp		0.957	0.030	0.96	0.527
Tra1	Omp		0.971	0.021	0.83	0.000
		Sod	0.985	0.025	1.00	1.000
Tra2	Omp		0.948	0.038	0.75	0.000
		Sod	0.956	0.034	0.93	0.399

[a] Ref1, Ref2: reference data sets; Tra1, Tra2: transfer data sets; all are derived from nucleotide sequences of bacterial *gap* loci. Omp: nucleotide sequences of bacterial *ompA* loci.

[b] Taxon designations as in Table 1.

[c] Significance estimated as the number of similarity coefficients generated by resampling methods that are smaller than that observed between the reference and transfer data sets.

erence data sets gave compatible matrices of identity coefficients, as did the two transfer data sets, which is expected since the data were derived from portions of the same genes (Table 2). However, when reference data sets were compared to transfer data sets, the similarity coefficient for the comparison was significantly smaller than the bootstrap distribution (P = 0.000 in 1000 bootstraps, Table 2). Reciprocal tests with resampling from the transfer data sets also yielded highly significant differences (data not shown). These results indicate that the taxonomic relationships implicit in the reference and transfer data sets are significantly different.

To determine the taxon to which this difference may be attributed, the analyses were repeated following the elimination of individual taxa from the data sets. Only the exclusion of taxon Sod, which was assigned alternate nucleotide sequences in the transfer data sets, results in reference and transfer data sets that describe compatible taxonomic relationships (P = 0.855, P = 0.880, Table 2). The exclusion of any other taxon did not yield data sets with compatible relationships. (As noted earlier, such multiple tests require appropriate adjustment of the significance levels.) The distribution of bootstrapped similarity coefficients for two cases of the Ref1/Tra2 simulation are shown in Figure 3. Considering all taxa (Figure 3A), it is clear that none of the bootstrapped similarity coefficients is smaller than the test statistic, which is more than 12 SD from the mean value of the bootstrap distribution. When taxon Sod is eliminated, the test statistic is not significantly different than the mean of the bootstrapped values. Comparison of the transfer data sets with a second independent set of nucleotide sequences (those from the bacterial *ompA* locus) also yielded statistically significant similarity coefficients (Table 2). As before, this difference was eliminated upon the removal of taxon Sod. Furthermore, the reference data sets for *gap* were not significantly different from the *ompA* sequences. These comparisons identify the transfer data sets as those resulting from the simulated horizontal genetic transfer event. Furthermore, the significance can be attributed to taxon Sod. Similar results were obtained when the simulation was repeated resampling the data using a jackknifing method (Wu 1986a) rather than the bootstrapping method (Table 3).

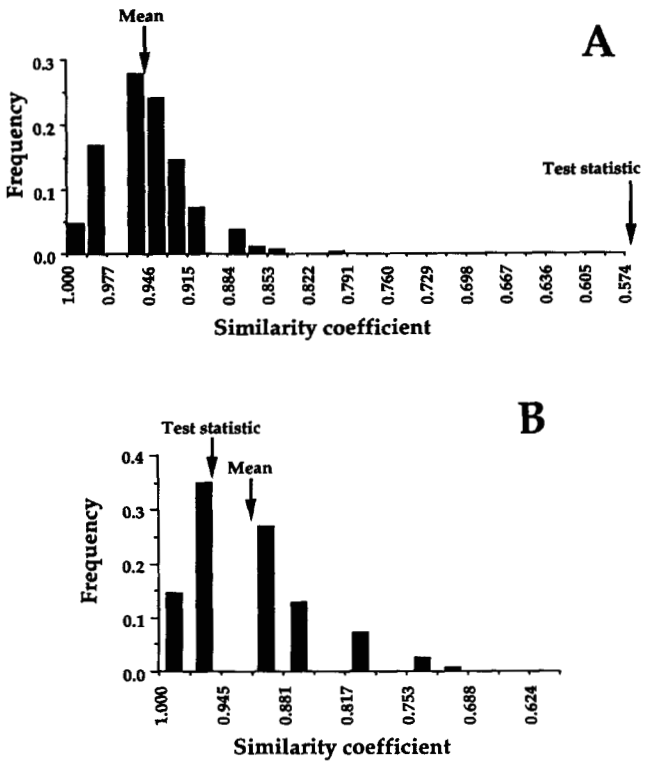

FIGURE 3.—The distribution of similarity coefficients for the Refl/Tra2 simulation (A) using all taxa and (B) eliminating taxon Sod.

We have also analyzed the data of MARUYAMA and HARTL (1991), which suggest that the transposable element *mariner* was transferred between a Drosophila taxon and an ancestor of another dipteran, *Zaprionus tuberculatus*. Portions of their data are summarized in Table 4. While nucleotide sequences encoding alcohol dehydrogenase (*Adh*) suggest that *Z. tuberculatus* is distantly related to all of the Drosophila taxa investigated, the *mariner* transposon resident in the genome of *Z. tuberculatus* is more closely related to the *mariner* transposons present in Drosophila taxa than was anticipated. Phylogenies approximating these relationships are shown in Figure 4. When compared using our methods, the *mariner* and *Adh* data sets described taxonomic relationships that are significantly different ($P < 0.01$, Table 5). (Reciprocal tests yielded comparable results.) However, when the *Z. tuberculatus* sequences were eliminated, the data sets described congruent relationships, as evident from the similarity coefficient of 1.0. The elimination of any other taxon did not yield this result. These results indicate that the inconsistency in taxonomic relationships described by the *Adh* and *mariner* sequences may be attributed to *Z. tuberculatus*. However, it is not clear from this test alone whether the discrepancy is due to transfer of a Drosophila *mariner* to Zaprionus or to an accelerated rate of evolution of *Adh* locus in *Z. tuberculatus*. Comparison of other genes from the same taxa, or *Adh* sequences from other Zaprionus species, would resolve the issue.

DISCUSSION

Several methods exist for discriminating between phylogenetic trees inferred from a single set of data (*e.g.*, FELSENSTEIN 1985; TEMPLETON 1983a). However, statistical frameworks have not been developed for directly testing the congruence of taxonomic relationships implicit in different data sets. Determining the robustness of a particular phylogeny to sampling errors in the data is an important issue. However, this problem may be separated from the question of whether the taxonomic relationships implicit in two data sets are significantly different, independent of what those relationships may be. By directly comparing the matrices of identity coefficients (or the matrices of ranks), the consistency of the taxonomic relationships may be analyzed without reference to phylogenetic inference. In the method we propose, the bootstrap is utilized to estimate the distribution of similarity coefficients resulting from the taxonomic relationships implicit in one data set. This distribution is utilized to assign a significance to comparisons between two data sets. Systematic elimination of individual taxa and reanalysis may reveal particular taxa with anomalous nucleotide sequences.

However, comparisons of additional genes, or algorithms that infer phylogenies, must also be employed to determine whether any inconsistencies are due to aberrations in evolutionary rates, homoplasy due to convergent evolution, or horizontal transfer of genetic material between distantly related taxa. If the elimination of anomolous sequences from the data sets also eliminates the significance of the difference, this may strengthen the case for horizontal genetic transfer. Homoplasy will result in the inability of typical algorithms to erect well defined phylogenetic trees, and the elimination of individual taxa from the data sets would not eliminate the inconsistencies. In some cases it may be difficult to distinguish between horizontal transfer of genetic material as opposed to the *ad hoc* hypothesis of an accelerated rate of evolution restricted to a small sequence of DNA evident in only a single lineage; for example, the discrepancy between the *Adh* and *mariner* data could also be explained by a much accelerated rate of evolution of the Zaprionus *Adh*, and excluding this possibility might require sequencing other genes.

Analyses of simulated data sets, which introduced an artificial horizontal transfer event into otherwise consistent data, provide evidence that the statistical method has the power to discriminate between sets of nucleotide sequences describing inconsistent taxonomic relationships. It is also clear from this simulation that the source of the aberrant nucleotide sequence can readily be identified. The statistical significance of the Drosophila *Adh* and *mariner* data supports the hypothesis of horizontal transfer of the *mariner* transposable element between these taxa.

TABLE 3

Data for simulated horizontal transfer event, 1000 jackknife repetitions

Jackknife data set[a]	Comparison data set	Taxon eliminated[b]	Distribution		Similarity coefficient	P^c
			μ	σ		
Ref1	Ref2		0.945	0.031	0.96	0.800
Ref2	Ref1		0.900	0.062	0.96	0.897
Tra1	Tra2		0.888	0.061	0.96	0.945
Tra2	Tra1		0.870	0.068	0.96	0.951
Ref1	Tra2		0.954	0.029	0.56	0.000
		Eco	0.941	0.045	0.65	0.000
		Sty	0.957	0.038	0.65	0.000
		Evu	0.968	0.033	0.60	0.000
		Kpn	0.960	0.039	0.40	0.000
		Sod	0.912	0.065	0.95	0.858
Ref2	Tra1		0.899	0.061	0.56	0.000
		Eco	0.915	0.067	0.60	0.000
		Sty	0.899	0.072	0.65	0.000
		Evu	0.935	0.066	0.60	0.000
		Kpn	0.920	0.071	0.45	0.000
		Sod	0.887	0.087	0.95	0.857

[a] Ref1, Ref2: reference data sets; Tra1, Tra2: transfer data sets.
[b] Taxon designations as in Table 1.
[c] Significance estimated as the number of similarity coefficients generated by resampling methods that are smaller than that observed between the reference and transfer data sets.

TABLE 4

Pairwise distances among dipteran taxa at the *Adh* locus (upper diagonal) and for *mariner* sequences (lower diagonal)

	Dma	Dsi	Dte	Dts	Dya	Ztu
Dma	—	0.012	0.051	0.095	0.049	0.184
Dsi	0.008	—	0.044	0.089	0.043	0.187
Dte	0.019	0.018	—	0.091	0.021	0.197
Dts	0.091	0.091	0.080	—	0.097	0.202
Dya	0.019	0.017	0.007	0.090	—	0.197
Ztu	0.027	0.026	0.023	0.105	0.023	—

Data from MARUYAMA and HARTL (1991). Taxon designations are: Dma, *Drosophila mauritiana*; Dsi, *D. simulans*; Dya, *D. yakuba*; Dte, *D. tessieri*; Dts, *D. tsacasi*; Ztu, *Zaprionus tuberculatus*. Entries are pairwise distances, calculated as the proportion of unchanged nucleotides between two taxa.

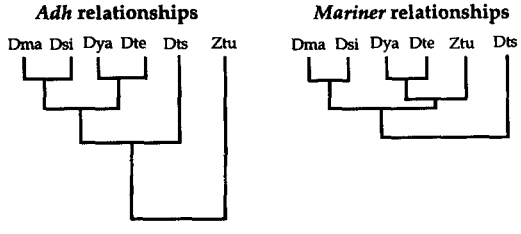

FIGURE 4.—Phylogenetic representation of the taxonomic relationships inferred from the *Adh* and *mariner* data sets (MARUYAMA and HARTL 1991). Taxon designations as in Table 4.

While horizontal transfer was suggested by visual inspection of the data, existing tests (TEMPLETON 1983a,b; Felsenstein 1988) did not attain statistical significance in support of this hypothesis MARUYAMA and HARTL, (1991).

As is the case in tests comparing phylogenies, significance values obtained from the bootstrap procedure must be adjusted to account for multiple tests. Not only are reciprocal tests performed, but in many cases partial data sets, comprising a subset of 5–7 taxa, may be examined to focus upon relationships among particular taxa. Careful selection of subsets, correction for multiple tests, and conservative interpretation of the data should aid in reducing the risk of spurious conclusions of horizontal genetic transfer.

The rationale for our proposed test is based on the assumption that the distribution of similarity coefficients generated by the bootstrap is valid. The justification for bootstrapping lies in the assumptions that all members of the data set (in our case nucleotide positions) are independent and identically distributed, and that the original data set is an unbiased sample of the underlying distribution. If the sample size is large, the second assumption is justified. Since polymorphisms in nucleotide sequences are usually not independent and identically distributed, one must propose both that the original data set is representative of the underlying distribution and that the resampling method preserves that structure. However, if this assumption is not valid, the procedure may be biased.

TABLE 5

Data for the Drosophila *Adh* and *mariner* comparisons, 1000 resampling repetitions

Resampled data set	Comparison data set	Taxa eliminated[a]	Distribution		Similarity coefficient	P^b
			μ	σ		
Bootstrap						
Adh	Mariner		0.961	0.027	0.87	0.010
		Dma	0.963	0.029	0.86	0.010
		Dsi	0.964	0.023	0.86	0.001
		Dte	0.977	0.027	0.86	0.004
		Dts	0.955	0.038	0.80	0.006
		Dya	0.958	0.036	0.86	0.029
		Ztu	0.946	0.042	1.00	1.000
Jackknife						
Adh	Mariner		0.960	0.027	0.87	0.004
		Dma	0.964	0.027	0.86	0.005
		Dsi	0.965	0.021	0.86	0.001
		Dte	0.978	0.026	0.86	0.005
		Dts	0.956	0.037	0.80	0.009
		Dya	0.959	0.037	0.86	0.030
		Ztu	0.942	0.041	1.00	1.000

[a] Taxon designations as in Table 4.

[b] Significance estimated as the number of similarity coefficients generated by resampling methods that are smaller than that observed between the *Adh* and *mariner* data sets.

It is not clear how such biases affect inference from nucleic acid sequences (FELSENSTEIN 1985, 1986).

To test the magnitude of resampling bias in our test, we employed the jackknife, a resampling procedure similar to the bootstrap (EFRON 1979; WU 1986a), but which is less sensitive to biases due to non-independent and identically distributed data sets. Whereas the bootstrap samples N values *with replacement* from a set of N observations, the jackknife selects r values *without replacement* from a set of N observations, where $0 < r < N$. In jackknifing a set of nucleotide positions, the variance may be estimated directly by selecting $r = 0.5 * N$ (FELSENSTEIN 1985, 1986; WU 1986a,b). Randomly selecting 50% of the data, without replacement, reduces the bias present in the bootstrap by selecting a smaller sample of the original data. To fully eliminate the bias, one must weight each jackknifed sample by the magnitude of its deviation from i.i.d. (WU 1986a), currently an intractable problem for nucleic acid sequences. However, when applied to the simulated data sets (Table 3) and the Drosophila *Adh* and *mariner* data sets (Table 5), the jackknife and the bootstrap yielded nearly identical results. Therefore, we conclude that the bootstrap accurately resamples the underlying distributions of nucleotide positions and does not yield significantly biased variance estimates.

In summary, we propose a framework for testing sets of nucleic acid sequences for inconsistencies resulting from horizontal transfer of genetic material between taxa. The method can detect inconsistencies in taxonomic relationships implicit in sets of nucleotide sequences without the *a priori* construction of phylogenetic trees. Using simulated data sets designed to represent horizontal transfer of genetic material, the method was successful in detecting aberrations in the data and in determining the taxon to which the aberrations could be attributed. Moreover, when applied to nucleotide sequences of the Drosophila *Adh* locus and the transposable element *mariner*, the method assigned high significance to a postulated horizontal transfer of *mariner* between ancestors of Drosophila and Zaprionus, whereas existing methods were unsuccessful in assigning significance to the differences in the phylogenies inferred from these sequences. The proposed statistical framework may also be applied to relationships inferred from restriction fragment length polymorphisms as well as amino acid sequences. In addition, a systematic examination of segments of genes may detect differences in the evolutionary histories of the segments resulting from intragenic recombination or exon shuffling. Programs for the PC environment for computing the statistics and distributions discussed in this paper are available upon request from J.G.L.

We are very grateful to JOSEPH FELSENSTEIN for his critical reading of the manuscript, for making important suggestions, and for pointing out ambiguities. We also thank STANLEY SAWYER ;or his interest and encouragement, and P. CAPY, J. CARULLI, D. KRANE, A. LARSON, H. OCHMAN for helpful discussions. This work was supported by grant 40322 from the National Institutes of Health.

LITERATURE CITED

BALDAUF, S. A., and J. D. PALMER, 1990 Evolutionary transfer of the chloroplast *tuf*A gene to the nucleus. Nature **344:** 262–265.

BANNISTER, J. V., and M. W. PARKER, 1985 The presence of a copper/zinc superoxide dismutase in the bacterium *Photobac-*

terium leignathi: a likely case of gene transfer from eukaryotes to prokaryotes. Proc. Natl. Acad. Sci. USA **82:** 149–152.

BRISSON-NOEL, A., M. ARTHUR and P. COURVALIN, 1988 Evidence for natural gene transfer from gram-positive cocci to *Escherichia coli*. J. Bacteriol. **170:** 1739–1745.

CARLSON, T. A., and B. K. CHELM, 1986 Apparent eukaryotic origin of glutamine synthetase II from the bacterium *Bradyrhizobium japonicum*. Nature **322:** 568–570.

DOOLITTLE, R. F., D. F. FENG, K. L. ANDERSON and M. R. ALBERRO, 1990 A naturally occurring horizontal gene transfer from a eukaryote to a prokaryote. J. Mol. Evol. **31:** 383–388.

DOW, M. M., and J. M. CHEVERUD, 1985 Comparison of distances matrices in studies of population structure and genetic microdifferentiation: Quadratic assignment. Am. J. Phys. Anthropol. **68:** 367–373.

DOWSON, C. G., A. HUTCHINSON, J. A. BRANNIGAN, R. C. GEORGE, D. HANSMAN, J. LIÑARES, A. TOMASZ, J. MAYNARD SMITH and B. G. SPRATT, 1989 Horizontal transfer of penicillin-binding protein genes in penicillin-resistant clinical isolates of *Streptococcus pneumoniae*. Proc. Natl. Acad. Sci. USA **86:** 8842–8846.

EFRON, B., 1979 Bootstrap methods: Another look at the jackknife. Ann. Statist. **7:** 1–26.

ELLIS, J., 1982 Promiscuous DNA - chloroplast genes inside plant mitochondria. Nature **299:** 678–679.

FARELLY, F., and R. A. BUTOW, 1983 Rearranged mitochondrial genes in the yeast nuclear genome. Nature **301:** 296–301.

FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:** 368–376.

FELSENSTEIN, J., 1985 Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:** 783–791.

FELSENSTEIN, J., 1986 Response to: jackknife, bootstrap and other resampling methods in regression analysis. Ann. Statist. **14:** 1304–1305.

FELSENSTEIN, J., 1988 Phylogenies from molecular sequences: inference and reliability. Annu. Rev. Genet. **22:** 521–565.

FIELD, K. G., G. J. OLSEN, D. J. LANE, S. J. GIOVANNONI, M. T. GHISELIN, E. C. RAFF, N. R. PACE and R. A. RAFF, 1988 Molecular phylogeny of the animal kingdom. Science **239:** 748–753.

FUKUDA, M., S. WAKASUGI, T. TSUZUKI, H. NOMIYAMA and K. SHIMADA, 1985 Mitochondrial DNA-like sequences in the human nuclear genome. Characterization and implications in the evolution of mitochondrial DNA. J. Mol. Biol. **186:** 257–266.

GELLISSEN, G., J. Y. BRADFIELD, B. N. WHITE and G. R. WYATT, 1983 Mitochondrial DNA sequences in the nuclear genome of a locust. Nature **301:** 631–634.

HILDEBRANDT, V., M. RAMEZANI-RAD, U. SWIDA, P. WREDE, S. GRZESIEK, M. PRIMKE and G. BÜLDT, 1989 Genetic transfer of the pigment bacteriorhodopsin into the eukaryote *Schizosaccaromyces pombe*. FEBS Lett. **243:** 137–140.

IWAASA, H., T. TAKAGI and K. SHIKAMA, 1989 Protozoan myoglobin from *Paramecium caudatum*. Its unusual amino acid sequence. J. Mol. Biol. **208:** 355–358.

JEFFS, P., and M. ASHBURNER, 1991 Processed pseudogenes in Drosophila. Proc. R. Soc. Lond. Ser. B **244:** 151–159.

KOLL, F., 1986 Does nuclear integration of mitochondrial sequences occur during senescence in Podospora? Nature **324:** 597–599.

LANDAN, G., G. COHEN Y. AHARONOWITZ, Y. SHUALI, D. GRAUR and D. SHIFFMAN, 1990 Evolution of isopenicillin N synthase genes may have involved horizontal gene transfer. Mol. Biol. Evol. **7:** 399–406.

LAWRENCE, J. G., H. OCHMAN and D. L. HARTL, 1991 Molecular and evolutionary relationships among enteric bacteria. J. Gen. Microbiol. **137:** 1911–1921.

LEUNISSEN, J. A. M., and W. W. DE JONG, 1986 Copper/zinc superoxide dismutase: how likely is gene transfer from ponyfish to *Photobacterium leignathi*? J. Mol. Evol. **23:** 250–258.

LIAUD, M.-F., D. X. ZHANG and R. CERFF, 1990 Differential intron loss and endosymbiotic transfer of chloroplast glyceraldehyde-3-phosphate dehydrogenase genes to the nucleus. Proc. Natl. Acad. Sci. USA **87:** 8918–8922.

MARÉCHAL-DROUARD, L., P. GUILLEMAUT, A. COSSET, M. ARBOGAST, F. WEBER, J.-H. WEIL and A. DIETRICH, 1990 Transfer RNAs of potato (*Solanum tuberosum*) mitochondria have different genetic origins. Nucleic Acids Res. **18:** 3689–3696.

MARUYAMA, K., and D. L. HARTL, 1991 Evidence for interspecific transfer of the transposable element mariner between *Drosophila* and *Zaprionus*. J. Mol. Evol. **33:** 514–524.

NEI, M., C. STEPHENS and N. SAITOU, 1985 Methods for computing the standard errors of branching points in an evolutionary tree and their application to molecular data from humans and apes. Mol. Biol. Evol. **2:** 66–85.

PEÑALVA, M. A., A. MOYA, J. DOPAZO and D. RAMON, 1990 Sequences of isopenicillin N synthetase genes suggest horizontal gene transfer from prokaryotes to eukaryotes. Proc. R. Soc. Lond. **241:** 164–169.

PENNY, D., and M. HENDY, 1986 Estimating the reliability of evolutionary trees. Mol. Biol. Evol. **3:** 403–417.

PERLER F., A. EFSTRATIADIS, P. LOMEDICO, W. GILBERT, R. KOLODNER and J. DODGSON, 1980 The evolution of genes: the chicken preproinsulin gene. Cell **20:** 555–566.

PLOS, K., S. I. HULL, B. R. LEVIN, I. ØRSKOV, F. ØRSKOV and C. SVANBORG-EDÉN, 1989 Distribution of the P-associated pilus (*pap*) region among *Escherichia coli* from natural sources: Evidence for horizontal gene transfer. Infect. Immun. **57:** 1604–1611.

SHATTERS, R. G., and M. L. KAHN, 1989 Glutamine synthetase II in Rhizobium: reexamination of the proposed horizontal transfer of DNA from eukaryotes to prokaryotes. J. Mol. Evol. **29:** 422–428.

SIEGEL, S., 1956 *Nonparametric Statistics for the Behavioral Sciences*. McGraw Hill, New York.

SNEATH, P. H. A., 1986 Estimating uncertainty in evolutionary trees from manhattan-distance triads. Syst. Zool. **35:** 470–488.

TEMPLETON, A. R., 1983a Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of humans and the apes. Evolution **37:** 221–244.

TEMPLETON, A. R., 1983b Convergent evolution and nonparametric inferences from restriction data and DNA sequences, pp 151–179 in *Statistical Analysis of DNA Sequence Data*, edited by B. S. WEIR. Marcel Dekker, New York.

TEMPLETON, A. R., 1985 The phylogeny of the Hominoid primates: a statistical analysis of DNA-DNA hybridization data. Mol. Biol. Evol. **2:** 420–433.

THOMAS, R. H., W. SCHAFFNER, A. C. WILSON and S. PÄÄBO, 1989 DNA phylogeny of the extinct marsupial wolf. Nature **340:** 465–467.

WAKABAYASHI, S., H. MATSUBARA, and D. A. WEBSTER, 1986 Primary sequence of a dimeric bacterial hemoglobin from vitreoscilla. Nature **322:** 481–483.

WOESE, C. R., 1987 Bacterial evolution. Microbiol. Rev. **51:** 221–271.

WRIGHT, R. M., and D. J. CUMMINGS, 1983 Integration of mitochondrial gene sequences within the nuclear genome during senescence in a fungus. Nature **302:** 86–88.

WU, C. F. J., 1986a Jackknife, bootstrap and other resampling methods in regression analysis. Ann. Statist. **14:** 1261–1295.

WU, C. F. J., 1986b Rejoinder to: jackknife, bootstrap and other resampling methods in regression analysis. Ann. Statist. **14:** 1343–1350.

Communicating editor: B. S. WEIR