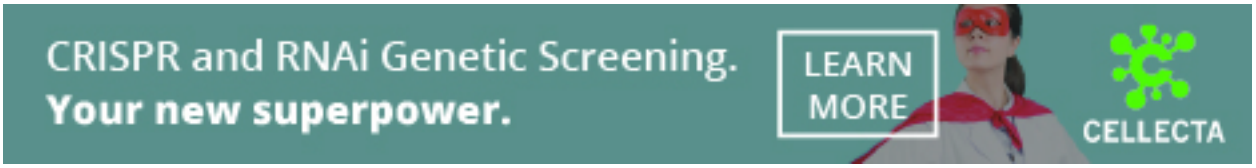
Computing genetic similarity coefficients from RAPD data: correcting for the effects of PCR artifacts caused by variation in experimental conditions.

W F Lamboy

Genome Res. 1994 4: 38-43



Computing Genetic Similarity Coefficients from RAPD Data: Correcting for the Effects of PCR Artifacts Caused by Variation in Experimental Conditions

Warren F. Lamboy

Department of Horticultural Sciences and U.S. Department of Agriculture–Agricultural Research Service (USDA–ARS), Plant Genetic Resources Unit, Cornell University, Geneva, New York 14456-0462

The production of informative random amplified polymorphic DNA (RAPD) markers using PCR and a single primer is often accompanied by the generation of artifactual (noninformative) bands as well. When RAPD data are used to compute genetic similarity coefficients, these artifacts (false positives, false negatives, or both) can cause large biases in the numerical values of the coefficients. As a result, some workers have been reluctant to use RAPD markers in the estimation of genetic similarities. Artifactual bands are of two types: those caused by variation in experimental conditions, and those caused by characteristics of the DNA to be amplified. A procedure is described that allows for correction of the bias caused by the first type of artifact, providing that replicate DNA samples have been extracted, amplified, and scored. The resulting data are used to obtain an estimate of the proportion of false-positive and false-negatives bands. These values are then used to correct the bias in the computed similarity coefficients. Two examples are given, one in which bias correction is critical to the results, and one in which it is less important. The maximum percent bias, computed from the estimated proportions of false positives and false negatives in the RAPD data set, is proposed as a criterion for determining whether bias correction of the similarity coefficients is required or not. Although all reasonable efforts should be made to optimize PCR protocols to eliminate artifactual bands, when this is not possible, the methods described allow RAPD markers to compute genetic similarities reliably and accurately, even when artifactual bands resulting from variation in experimental conditions are present.

In a previous paper,[1] I examined the properties of and relationships among three coefficients that measure genetic similarities using data on random amplified polymorphic DNA (RAPD) markers.[2,3] The coefficients considered were the simple matching coefficient,[4] Jaccard's coefficient,[4,5] and Nei and Li's coefficient.[6]

Amplification of DNA fragments using PCR with a single primer not only produces informative bands but also artifactual bands, both false positive and false negative. The many different sources of artifactual bands have been discussed.[1] These artifacts cause bias (underestimation or overestimation) in the computed values of all three similarity coefficients, but Nei and Li's coefficient is affected less by artifacts than are the other two. For this and other reasons, Nei and Li's coefficient was recommended[1] for use in the routine computation of genetic similarities using RAPD data, when correction for artifacts is either not possible or not desirable.

Artifacts may be caused by either variation in experimental conditions (DNA purity, Mg^{2+} concentration, type of thermocycler, etc.) or by characteristics associated with the DNA itself (repetitive sequences, genetic background, in vitro recombinants, etc.).[1] Only the former type of artifactual bands are considered in this manuscript, for they can be detected using the procedures described. Detection of the latter type of artifact is beyond the scope of this paper, as complex analysis or ancillary information is necessary to identify these bands.[7–10]

Every effort should be made to optimize PCR protocols to minimize the number of artifactual bands that are generated. In some instances, however, optimization may be prohibitively expensive or time-consuming. In this paper I show that it is still possible to estimate genetic similarities accurately under these circumstances. Through the use of replicate observations, the levels of false positives and false negatives in the data are estimated. These estimates are used to correct for the bias in the computed values of the similarity coefficients. Bias correction results in more accurate estimates of similarities among samples, allows for the comparison of similarity estimates from different data sets, and increases the accuracy of the relationships determined using phenetic (cluster) analyses and similar analytical methods.

The purpose of this paper is to explain how replicate observations are used in computing genetic similarity estimates that are corrected for the presence of artifactual bands (false positives, false negatives, or both). Two examples are

4:38–43©1994 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/94 $5.00

presented, one in which correcting for artifactual bands causes great differences in the results obtained, and the second where the differences are minor. By combining these results with those obtained previously,[1] the conditions under which correction for artifactual bands is worth the additional effort and expense are determined.

MATERIALS AND METHODS

Assume that there are n different DNA samples for which pairwise genetic similarities using RAPD data are to be computed. To avoid excess numbers of subscripts and to simplify the presentation, consider the computation for just two samples. Call them samples i and j. Assume further that the two samples share some of the same positive bands and lack some of the same negative bands. Let the total proportion of shared positive and negative bands be equal to s. Let p represent the proportion of the shared bands that are shared because both samples i and j possess them. These are the shared positive bands. Note that, in general, the value for p will be different for every pair of samples compared. Then $1-p$ is the proportion of bands shared because the samples lack the same bands. These are the shared negative bands.

The proportion of bands that the samples do not share is $1-s$. Because these bands are unshared, any specific band must be present in either sample i or sample j, but not both. Let r be the proportion of unshared bands that are present in sample i. Then r also will represent the proportion of unshared bands that are absent in sample j. The proportion of unshared bands that are absent in sample i will be $1-r$, and the proportion of unshared bands that are present in sample j is also $1-r$.

Let f_p be the probability that a band that should not be present in a sample is present because of experimental error; it is the probability of a false positive. Let f_n be the probability that a band that should be present in a sample is absent because of experimental error; it is the probability of a false negative.

Assume that there are m_i replicate RAPD runs of sample i and m_j replicate RAPD sample runs of sample j, where m_i, $m_j \geq 1$. For sample runs to be true replicates, which account for all sources of variability in an experiment, every step of the procedure, from DNA extraction to the running of amplified products on a gel, must be performed independently for all replicates of a sample. If not all the factors can be replicated, for example, because they are too expensive or too time consuming, then the full range of artifactual bands probably will not be generated. While this situation is less than ideal, it is certainly better to account for some sources of variability, with less than ideal replicates, than to account for none of them by running no replicates at all.

Finally, some terminology is needed. "True value," means the value that would result if there were no false positives and no false negatives. "Estimated value," or "estimate," designates the value that results when the data containing false positives and false negatives is used in the usual way to compute values for the coefficients. "Corrected value" refers to the values of the coefficients computed in a way that corrects for the false positives and false negatives that are present.

RESULTS AND DISCUSSION

Estimating the Total Similarity, s, and Proportion of Similarity Caused by Shared Positive Bands, p, When There are Replicate Runs

When there are replicate runs, it is easiest to estimate s as the sum of the estimated similarity caused by shared positive bands, $s \times p$ plus the estimated similarity caused by shared negative bands, $s \times (1-p)$. Table 1 shows the results of computation of these values for one pair of samples.

With replicate observations, the first step in computing $s \times p$ is to determine the proportion of times the band was present in each sample. For each band, the proportion of times it was present in the replicates of sample i is multiplied by the proportion of times it was present in the replicates of sample j. The products are summed over all bands and divided by the total number of bands. The result is an estimate of $s \times p$.

The computation of the value for $s \times (1-p)$ is analogous. For each band and each sample, the proportion of times the band was absent (this is just 1 minus the proportion of times the band was present) is determined. Then for each band, the proportion of times the band was absent in the replicates of sample i multiplied by the proportion of times it was absent in the replicates of sample j is found. The products are summed over all bands and divided by the total number of bands, giving the value for $s \times (1-p)$.

To find the estimate for s, simply add the estimates of $s \times p$ and $s \times (1-p)$. The estimate for p can be obtained by dividing the estimate of $s \times p$ by the estimate for s.

Estimating f_p and f_n

The values of f_p and f_n are estimated from the replicates runs for each sample. Recall that a false positive is defined to be a band that ought not to be present but is; a false-negative band is a band that ought to be present but is not. Thus, any band that is present in one replicate, but not in another, can be viewed as either a false positive or a false negative, depending on which of the replicates shows the band's "true" state. The pro-

TABLE 1 Computation of Estimates of s and p from Two RAPD Samples, Having Two and Three Replicates, Respectively

Band	Replicates from sample i		Proportion of bands present	Proportion of bands absent	Replicates from sample j			Proportion of bands present	Proportion of bands absent
1	1	1	1.00	0.00	1	0	1	0.67	0.33
2	0	1	0.50	0.50	0	0	0	0.00	1.00
3	1	1	1.00	0.00	0	0	1	0.33	0.67
4	0	0	0.00	1.00	0	1	0	0.33	0.67
5	1	1	1.00	0.00	1	1	1	1.00	0.00

(0) Band absent; (1) band present.
Estimate of $s \times p = 0.400$; estimate of $s \times (1-p) = 0.233$; solving for s gives, $s = 0.633$; solving for p gives $p = 0.633$.
In this example the estimate of s = the estimate of p, is a coincidence. Generally, the two estimates will not be equal.

portion of false positives is the proportion of bands that are present in a replicate but absent in another replicate assumed to represent the true state. Because all replicates are equally likely to show the true situation, the overall proportion is computed considering each replicate in turn as true and averaging the results over all pairs of replicates in a sample. Proportions of false negatives can be computed analogously. After the computations are performed for each sample separately, they are averaged over all samples.

To compute estimates of f_p and f_n from a single sample, for example, sample i, the procedure is as follows. First, make all possible ordered pairs of replicates. If there are m_i replicates of sample i, then there will be $m_i \times (m_i - 1)$ different ordered pairs of replicates. In each case, the first replicate in a pair is taken as true. To compute the proportion of false positives, count the number of bands that are present in the second replicate but absent in the first replicate and divide that number by the number of bands present in the second replicate. This supplies an estimate of the proportion of bands observed to be present, which, based on other information (namely the bands in the first replicate), ought not to be present. To compute the proportion of false negatives, find the number of bands that are absent in the second replicate but present in the first replicate and divide that number by the number of bands absent in the second replicate. This supplies an estimate of the proportion of bands observed to be absent, which, based on other information (the bands in the first replicate), ought to be present. Each ordered pair of replicates is examined in the same way, and each pair supplies one estimate of f_p and one estimate of f_n, with every inconsistent band being treated as *both* a false positive and a false negative to obtain these initial estimates (this "double counting" will be adjusted for below). Adding the estimates of the proportions of false positives and false negatives and dividing by $m_i \times (m_i - 1)$ gives the estimates of the values for f_p and f_n from sample i. Summing the values from all samples and dividing by n gives the initial estimates of f_p and f_n from all n samples. This method of computation gives equal weight to each sample, regardless of how many replicates were run for it.

These initial estimates of f_p and f_n represent the *maximum* possible proportions of false positives and false negatives that could be present in the data, because each inconsistent band is treated as both a false positive and a false negative. In effect, each band is counted twice. The initial estimates need to be adjusted based either on ancillary information or on other assumptions that the experimenter is willing to make, so that each inconsistent band is counted only once. There are four possible ways to do this. (1) If the researcher believes that only false positives are likely to occur in the data, then the initial estimate for f_p should be used as the final value, and f_n should be set $=0$. (2) If, on the other hand, only false negatives are believed to be present, $f_p = 0$ should be set and the initial estimate of f_n used as the final value. (3) If there is evidence that false positives and false negatives are equally frequent, the initial estimates of f_p and f_n should be averaged, and the final value of both f_p and f_n should be set to equal one-half of that average. (4) If equal weight is to be given to the estimates of false positives and false negatives but there is no evidence to suggest that they occur with equal frequency, then f_p should be set to equal one-half the estimated initial value of f_p and f_n should be set to equal one-half the estimated initial value of f_n. Each band is counted only once for each of these alternatives.

Using the data from Table 2 as an example, after computing final estimates of f_p and f_n using the four alternative methods, one has values of (f_p, f_n) equal to (0.271, 0.000), (0.000, 0.306), (0.144, 0.144), or (0.136, 0.153), respectively. Careful selection of the method for the final estimates of f_p and f_n is essential, as the values computed for f_p and f_n by the different methods can differ greatly. In the absence of ancillary information indicating otherwise, I recommend use of the fourth method for estimating the proportions of the two types of artifactual bands, as it assumes the least a priori knowledge about the relative frequencies of false positives and false negatives.

Using the notation defined above, it was shown[1] that the values for the simple matching coefficient, (SMC), Jaccard's coefficient (J), and Nei and Li's coefficient (NL), are given by the following equations:

$$\mathrm{SMC} = sp + s(1 - p) \qquad (1)$$

TABLE 2 Computation of Estimates of Maximum Possible Values of f_p and f_n From Samples i and j of Table 1

"True replicate"	Compared with replicate	Estimate of f_p	Estimate of f_n
	Sample i		
1	2	0.250	0.000
2	1	0.000	0.500
Average (maximum)		0.125	0.250
	Sample j		
1	2	0.500	0.333
1	3	0.333	0.000
2	1	0.500	0.333
2	3	0.667	0.500
3	1	0.000	0.333
3	2	0.500	0.667
Average (maximum)		0.417	0.361
Average (maximum, both samples)		0.271	0.306

In an actual data set, estimates of f_p and f_n would be obtained using data from all samples (see text).

$$J = \frac{sp}{sp + (1 - s)r + (1 - s)(1 - r)} \qquad (2)$$

$$J = \frac{sp}{sp + (1 - s)}$$

$$NL = sp/[sp + (1 - s)r + sp + (1 - s)(1 - r)]/2 = \frac{2sp}{2sp + (1 - s)} \qquad (3)$$

It was also shown that the estimates produced for SMC, J, and NL (when there were artifactual bands present) were as follows:

$$est(SMC) = f_n + f_p - 2f_nf_p + (-1 + f_n + f_p)(-1 + 2f_p + 2pf_n - 2pf_p)s \qquad (4)$$

$$est(J) = [p(f_n + f_p - 2f_nf_p) + sp(-1 + f_n + f_p)(-1 + 2f_p + 2pf_n - 2pf_p)]/[1 - f_n - f_p + 2f_nf_p + p(f_n + f_p - 2f_nf_p) + s(-1 + f_n + f_p)(-1 + p)(-1 + 2f_p + 2pf_n - 2pf_p)] \qquad (5)$$

$$est(NL) = [2p(f_n + f_p - 2f_nf_p) + 2sp(-1 + f_n + f_p)(-1 + 2f_p + 2pf_n - 2pf_p)]/[1 - f_n - f_p + 2f_nf_p + 2p(f_n + f_p - 2f_nf_p) + s(-1 + f_n + f_p)(-1 + 2p)(-1 + 2f_p + 2pf_n - 2pf_p)] \quad (6)$$

After est(SMC), est(J), and est(NL) are computed for a set of data, equation 4, 5, or 6, respectively, can be used to solve for the corrected value for s, cor(s), provided p, f_n, and f_p have been estimated as described above. The solutions for cor(s) are given in equations 7–9. All three give the same numerical value for cor(s). Use of SMC gives:

$$cors(s) = [-a + f_n + f_p - 2f_nf_p]/[(-1 + f_n + f_p)(1 - 2f_p - 2pf_n + 2pf_p)] \quad (7)$$

where $a = est(SMC)$.

J gives:

$$cor(s) = [b - bf_n - bf_p + 2bf_nf_p + p(-f_n + bf_n - f_p + bf_p + 2f_nf_p - 2bf_nf_p)]/[(-1 + f_n + f_p)(-b - p + bp)(1 - 2f_p - 2pf_n + 2pf_p)] \quad (8)$$

where $b = est(J)$.

NL gives:

$$cor(s) = [c - cf_n - cf_p + 2cf_nf_p + p(-2f_n + 2cf_n - 2f_p + 2cf_p + 4f_nf_p - 4cf_nf_p)]/[(-1 + f_n + f_p)(-c - 2p + 2cp)(1 - 2f_p - 2pf_n + 2pf_p)] \quad (9)$$

where $c = est(NL)$. Once computed, the value of cor(s) can be substituted into equations 1–3 to obtain corrected values for SMC, J, or NL. The corrected values are then available for use in phenetic (cluster) analyses, for comparisons among similarity coefficients computed for other samples or by other methods, and so forth. Although these equations appear to be complex, once the proper quantities have been estimated, they are straightforward algebra and can be computed readily.

EXAMPLES AND DISCUSSION

Example 1 in Table 3 shows that there sometimes may be dramatic differences among the uncorrected and corrected values of the similarity coefficients. The values of Nei and Li's coefficient were computed for seven samples each having two replicates (raw data not shown). The top matrix shows the uncorrected values, and the bottom matrix shows the corrected values of the coefficients. The estimates of f_p and f_n computed from the raw data were 0.115 and 0.172, respectively, weighting the two types of artifacts equally but not assuming they occur with equal frequency. Differences between uncorrected and corrected values in Table 3 range from 0.281 for the similarity between samples 6 and 7, to −0.142 for samples 3 and 5.

TABLE 3 Uncorrected (Top Matrix) and Corrected (Bottom Matrix) Values for Nei and Li's Coefficient when $f_p = 0.115$ and $f_n = 0.172$

	1	2	3	4	5	6	7
1	1.000	0.476	0.389	0.500	0.429	0.000	0.630
2	0.476	1.000	0.524	0.435	0.667	0.267	0.667
3	0.389	0.524	1.000	0.400	0.667	0.167	0.556
4	0.500	0.435	0.400	1.000	0.435	0.429	0.690
5	0.429	0.667	0.667	0.435	1.000	0.333	0.733
6	0.000	0.267	0.167	0.429	0.333	1.000	0.286
7	0.630	0.667	0.556	0.690	0.733	0.286	1.000
	1	2	3	4	5	6	7
1	1.000	0.427	0.333	0.497	0.328	0.000	0.654
2	0.427	1.000	0.525	0.273	0.768	0.215	0.702
3	0.333	0.525	1.000	0.300	0.809	0.149	0.472
4	0.497	0.273	0.300	1.000	0.273	0.500	0.762
5	0.328	0.768	0.809	0.273	1.000	0.314	0.842
6	0.000	0.215	0.149	0.500	0.314	1.000	0.005
7	0.654	0.702	0.472	0.762	0.842	0.005	1.000

When the similarity data for example 1 was used in a UPGMA (unweighted pair-group method using arithmetic averages)[4] cluster analysis carried out by the program NTSYS,[11] different results were obtained from the use of uncorrected and the corrected coefficients. Figure 1 shows the result using uncorrected coefficients, and Figure 2 shows the results using the corrected coefficients. The groups that are the same in the two phenograms are the clusters containing samples [5,7], [2,5,7], and [2,3,5,7]. Groups involving samples 1, 4, and 6 are different on the two phenograms. In this example, bias correction, or lack of it, significantly affected the results obtained.

In contrast, example 2 in Table 4 shows little difference between the uncorrected and corrected values of Nei and Li's similarity coefficient. The estimates of f_p and f_n here are 0.00245 and 0.00116, respectively. All of the differences among uncorrected and corrected coefficients are ≤0.005 in magnitude.

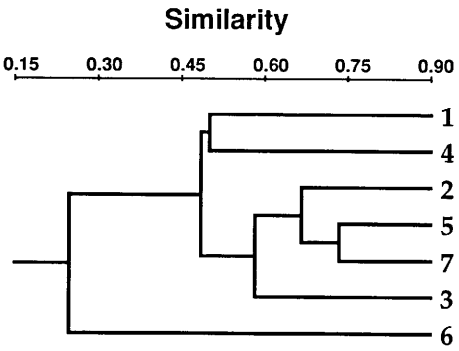


FIGURE 1 Results of UPGMA cluster analysis using uncorrected values of Nei and Li's similarity coefficient of similarity and data from the to matrix in in Table 3.

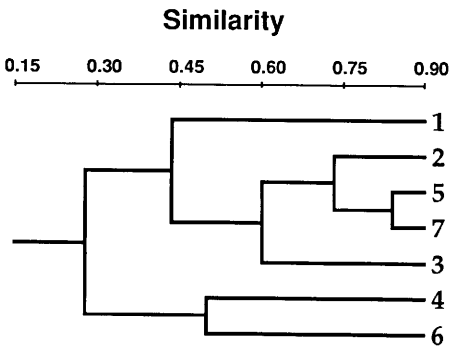


FIGURE 2 Results of UPGMA cluster analysis using corrected values of Nei and Li's coefficient of similarity and data from the top matrix in Table 3.

Figures 3 and 4 show the phenograms produced by a UPGMA cluster analysis using Nei and Li's coefficient. The topologies of the phenograms are identical, although there are some small differences in the lengths of some of the branches. Whether coefficients were corrected or not had little effect on the results of the cluster analysis.

The two examples show that depending on the levels of false positives and false negatives, correction of the similarity coefficients may or may not make a substantial difference in the results. How can it be determined whether bias correction is needed? One approach to answering this question is as follows.

First, it has been shown[1] that the greatest magnitudes of percent bias occurred at the low values (<0.10) for s. As s increased, the magnitude of the percent bias decreased to a minimum at $s=0.50$ and then began to increase again as s approached 0.90. The tables in a previous paper[1] show that for values of $s<0.10$ or $s>0.90$ the percent bias can be computed approximately as $(f_p+f_n/s\times 100\%$. The tables also show that the percent bias for values of s between 0.10 and 0.90 will be less than the percent bias outside that range. Thus, a conservative estimate of the upper limit on the maximum possible percent bias for ALL values of s is $(f_p + f_n)/s \times 100\%$. For instance, in example 1 (Table 3) the maximum percent bias for $s=0.10$ is $\sim(.287/.100)\times 100\%=287\%$, and for $s=0.90$ is $\sim 287/.900\times 100\%=32\%$. In example 2, the corresponding values are 4% and 0.5%.

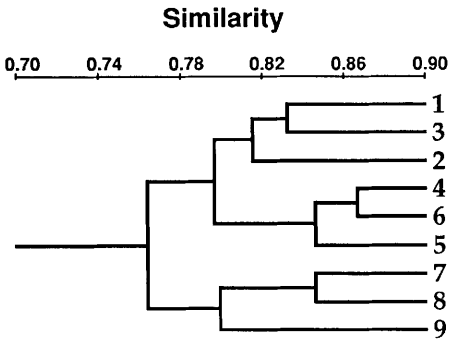


FIGURE 3 Results of UPGMA cluster analysis using uncorrected values of Nei and Li's coefficient of similarity and data from the bottom matrix in Table 4.

These estimated values of the maximum percent bias can be used as guidelines for determining whether the correction needs to be applied to the estimated values of the similarity coefficients in any particular situation. On the basis of empirical studies of these coefficients, I suggest that when the percent bias is greater than two times the difference between the two similarities in a data set that differ the least, it is likely that correcting the coefficients for artifacts will have a noticeable effect on the results of a cluster analysis. If the maximum percent bias is less than two times the difference between the two similarities that differ the least, correcting the coefficients probably will not change relationships found in a cluster analysis significantly, although it still would provide more accurate estimates of similarity. If no correction for artifacts is possible, then use of Nei and Li's similarity coefficient should be considered, for that has been shown[1] to be affected least by false positives and false negatives.

Not all of the DNA samples under study need to be replicated to obtain estimates of f_p and f_n as long as those that are selected for replication are chosen randomly. Because the approximate variance of the estimate of, for example, f_p, is given by $f_p\times(1-f_p)/(n_r\times n_b)$, where n_r is the number of samples for which replicates were taken, and n_b is the number of RAPD bands analyzed (the estimated variance of f_n is computed similarly by substituting f_n for f_p), one can determine the number of samples that need to be replicated to ensure that the estimated standard deviation of the estimate of f_p (or f_n) is less than some specified value.

It should be emphasized that if the exact values of the similarity coefficients are sought, then artifacts always should be corrected for. If the unbiased values of the coefficients are not of intrinsic interest but will be used in phenetic analyses or other similar procedures, then if there is low bias, correction probably is not necessary. When percent bias is large, however, the coefficients always should be corrected for artifactual bands, regardless of their eventual use.

TABLE 4 Uncorrected (Top Matrix) and Corrected (Bottom Matrix) Values for Nei and Li's Coefficient when f_p = 0.00245 and f_n = 0.00163

	1	2	3	4	5	6	7	8	9
1	1.000	0.826	0.832	0.802	0.787	0.754	0.789	0.699	0.754
2	0.826	1.000	0.805	0.820	0.771	0.784	0.760	0.704	0.760
3	0.832	0.805	1.000	0.854	0.805	0.795	0.843	0.805	0.759
4	0.802	0.820	0.854	1.000	0.851	0.867	0.822	0.738	0.745
5	0.787	0.771	0.805	0.851	1.000	0.841	0.750	0.683	0.761
6	0.754	0.784	0.795	0.867	0.841	1.000	0.786	0.795	0.795
7	0.789	0.760	0.843	0.822	0.750	0.786	1.000	0.846	0.818
8	0.699	0.704	0.805	0.738	0.683	0.795	0.846	1.000	0.780
9	0.754	0.760	0.759	0.745	0.761	0.795	0.818	0.780	1.000
	1	2	3	4	5	6	7	8	9
1	1.000	0.830	0.836	0.805	0.790	0.757	0.792	0.702	0.757
2	0.830	1.000	0.809	0.823	0.774	0.787	0.764	0.708	0.763
3	0.836	0.809	1.000	0.858	0.808	0.799	0.848	0.810	0.762
4	0.805	0.823	0.858	1.000	0.855	0.871	0.826	0.741	0.747
5	0.790	0.774	0.808	0.855	1.000	0.845	0.753	0.686	0.764
6	0.757	0.787	0.799	0.871	0.845	1.000	0.789	0.799	0.799
7	0.792	0.764	0.848	0.826	0.753	0.789	1.000	0.851	0.822
8	0.702	0.708	0.810	0.741	0.686	0.799	0.851	1.000	0.784
9	0.757	0.763	0.762	0.747	0.764	0.799	0.822	0.784	1.000

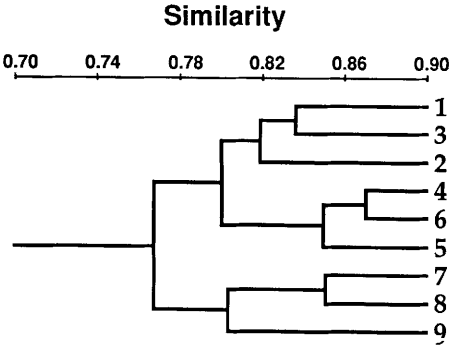


FIGURE 4 Results of UPGMA cluster analysis using corrected values of Nei and Li's coefficient of similarity and data from the bottom matrix in Table 4.

The best solution to the problem of artifactual RAPD bands caused by experimental error is to optimize DNA extraction and amplification protocols so that bands are consistent across replicates. When that is not possible, however, artifactual band variation still can be corrected for by use of the equations given above. With continuing improvement in amplification protocols and techniques and the ability to correct for false positives and false negatives, RAPDs now can be used for reliably computing genetic similarities and for accurately determining relationships among organisms even when artifactual bands caused by variation in experimental conditions are present.

ACKNOWLEDGMENTS

This manuscript benefited from discussions with James McFerson, Anne Westman, and Muhammad Lodhi. The comments and helpful criticisms of Robert Bernatzky, Bruce Reisch, Ray Schnell, Mark Sorrells, and an anonymous reviewer were greatly appreciated. Kathy Ronning's careful reading and insightful questions helped to improve the manuscript significantly.

REFERENCES

1. Lamboy, W. 1994. Computing genetic similarity coefficients from RAPD data: The effects of PCR artifacts. *PCR Methods Applic.* (this issue).
2. Welsh, J. and M. McClelland. 1990. Fingerprinting genomes using PCR with arbitrary primers. *Nucleic Acids Res.* **18:** 213—7218.
3. Williams, J.G.K., A.R. Kubelik, K.J. Livak, J.A. Rafalski, and S.V. Tingey. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.* **18:** 6531–6535.
4. Sneath, P.H.A. and R.R. Sokal. 1973. *Numerical taxonomy*. W.H. Freeman and Company, San Francisco, CA.
5. Jaccard, P. 1908. Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaudoise. Sci. Nat.* **44:** 223–270.
6. Nei, M. and W.-H. Li. 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci.* **76:** 5269–5273.
7. Jansen, R. and F.D. Ledley. 1990. Disruption of phase during PCR amplification and cloning of heterozygous target sequences. *Nucleic Acids Res.* **18:** 5153–5156.
8. Hunt, G.J. and R.E. Page Jr. 1992. Patterns of inheritance with RAPD molecular markers reveal novel types of polymorphism in the honey bee. *Theor. Appl. Genet.* **85:** 15–20.
9. Riedy, M.F., W.J. Hamilton III, and C.F. Aquadro. 1992. Excess of non-parental bands in offspring from known primate pedigrees assayed using RAPD PCR. *Nucleic Acids Res.* **20:** 918.
10. Ayliffe, M.A., G.J. Lawrence, J.G. Ellis, and A.J. Pryor. 1994. Heteroduplex molecules formed between allelic sequences cause nonparental RAPD bands. *Nucleic Acids Res.* **22:** 1632–1636.
11. Rohlf, F.J. 1988. NTSYS-pc, Numerical taxonomy and multivariate analysis dystem. Exeter Software, Setauket, New York.

Received April 15, 1994; accepted in revised form June 15, 1994.