Amer J Hum Genet 25:422–432, 1973

Topology of Kinship in Micronesia

N. E. Morton[1] and J. M. Lalouel[1]

All the information about kinship of *n* populations is encoded in a square, symmetric matrix Φ whose element φ_{ij} bears a simple relation to the probability that a random gene in population *i* be identical by descent with a random gene in population *j*. When estimated from phenotype bioassay as conditional kinship, φ_{ij} is an estimate of the correlation between a random gamete in *i* and a random gamete in *j* relative to the contemporary gene frequencies. Such a matrix is given in Morton and Lalouel [1, table 4], who also presented corresponding estimates from anthropometrics [1, table 5]. These biological indicators may be compared with data on frequency of shared cognates [2, table 5]. The object of this paper is to analyze the relationships implicit in these matrices by mapping them in a reduced dimensionality, using eigenvectorial and dendrogram representations, without making unnecessary, untestable, or implausible phyletic assumptions.

GEOGRAPHY

Coordinates are usually given in degrees and minutes. Let

$$\begin{aligned} x' &= \pm\,(x + m_x/60), \\ y' &= \pm\,(y + m_y/60), \end{aligned} \tag{1}$$

where x' is longitude in decimal notation, x is degrees of longitude, and m_x is minutes of longitude, the sign being positive for the Eastern Hemisphere. Similarly, y' is decimal latitude, y is degree of latitude, and m_y is minutes of latitude, the sign being positive north of the equator. Of course, if s_x, s_y are seconds, the quantities $s_x/3{,}600$ and $s_y/3{,}600$ are added to x, y, respectively, but such precision is rarely required. Denote the means over the array of *n* populations by $\overline{x'}$, $\overline{y'}$. Using plane trigonometry on the Hayford spheroid, the coordinates in kilometers as deviations from the means are

$$\begin{aligned} X &= (x' - \overline{x'})\,(111.4175\cos B - 0.0940\cos 3B + 0.0002\cos 5B) \\ Y &= (y' - \overline{y'})\,(111.1363 - 0.5623\cos 2B + 0.0011\cos 4B), \end{aligned} \tag{2}$$

where the middle latitude is $B = (y' + \overline{y'})/2$ [3]. Spherical trigonometry gives

Received April 25, 1972; revised February 26, 1973.

Population Genetics Laboratory paper no. 96. This work was supported by grant GM 17173 from the U.S. National Institutes of Health.

[1] Population Genetics Laboratory, University of Hawaii, Honolulu, Hawaii 96822.

© 1973 by the American Society of Human Genetics. All rights reserved.

great circle distances, but this refinement is unnecessary for the distances that are important in population structure.

The geography of 14 Micronesian samples is shown in figure 1, indicating coordinates as deviations in kilometers from the central point (153°40′ E, 7°20′ N),

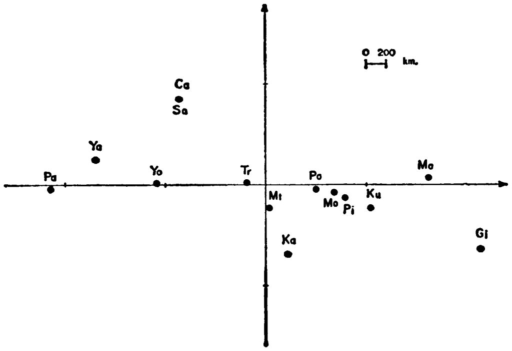

FIG. 1.—Geography of 14 Micronesian samples

and assigning Carolinians on Saipan to the place of residence, not origin [1]. Only the four samples in the Ponape district are sufficiently close for local migration among defined populations to be appreciable. There is, of course, approximation in assuming the Marshallese and Gilbertese panmictic, although all evidence testifies to frequent internal migration in these archipelagoes [4, 5]. Because the space is a narrow ellipse, we might expect any representation of population structure to emphasize east-west differentiation. Kapingamarangi as a Polynesian outlier has a special status not reflected by its geographic location.

EIGENVECTORS OF KINSHIP

To examine the phenetic relationships implied by a kinship matrix, we must reduce the dimensionality of the space of observations. In practice we consider two-dimensional representations, which are readily comprehended and correspond to the geographic dimensionality.

A well-known procedure is called principal component extraction when applied to a matrix of the type $A = XX'$, which is positive semi-definite (i.e., without negative eigenvalues). For conditional kinship (relative to contemporary gene frequencies), some of the smaller eigenvalues may be negative, and so we use another theory which assures that the absolute value of the sum of squared distances between points for any vector k is proportional to the absolute eigenvalue $|\lambda_k|$, providing the matrix has been centroid adjusted, so that

$$a_{ij}' = a_{ij} - a_{i.} - a_{.j} + a_{..}, \qquad (3)$$

where a_{ij} is an element before adjustment and $a_{i.}$, $a_{.j}$, and $a_{..}$ are the corresponding row, column, and overall means, respectively [6]. Two coefficients are computed that measure the effect of dimensionality reduction: a product-moment correlation

R between the centroid-adjusted matrix and its two-dimensional approximation and the fraction

$$f(\lambda) = (|\lambda_1| + |\lambda_2|) \,/ \sum_{i=1}^{n} |\lambda_i|$$

of the total absolute dispersion extracted by those two vectors.

In figure 2, these representations have been translated and rotated to maximum congruence with geography [6], using the program MATFIT to solve by least squares, $B = AT + J\gamma' + E$, following [7]. The B and A are $(n \times 2)$ matrices of the geographical coordinates and leading eigenvectors of the adjusted matrix, respectively; T, a (2×2) orthogonal matrix; J, an $(n \times 1)$ unit vector; γ, a (2×1) translation vector; and E, an $(n \times 2)$ matrix of residuals. This process leaves the configuration invariant. The closed figures correspond to clustering levels such that mean hybridity within clusters is less than .01 or .02.

Phenotype bioassay gives a topology emphasizing east-west differentiation. Kapingamarangi is clearly differentiated from the rest. Pingelap and the Marshalls are displaced toward the west in curious association with Palau and Yap, whereas most of eastern Micronesia forms one cluster. Anthropometrics also separate Kapingamarangi and emphasize east-west differentiation. Cognate frequencies differentiate the Heonesian subfamily of central and eastern Micronesia from the three western populations, which themselves show little affinity.

DENDROLOGY

In eigenvectorial topology, phenotypically similar populations tend to be located close together in a plane. There is another type of graph, called a "tree," in which dissimilarity is indicated by distance along one axis, the orthogonal axis serving merely to space the populations uniformly, so that the logical form of the tree is unchanged by rotation of any branch, as $(AB)(C) = (BA)(C) = (C)(AB) = (C)(BA)$. A phenetic tree, or "dendrogram," is based entirely on phenotypic dissimilarity. An appropriate parameter with a range from zero to one is hybridity [8], defined as the excess of heterozygosity in an F_1 generation compared with an F_2 between the pair of populations, or

$$\theta_{ij} = \frac{\varphi_{ii} + \varphi_{jj} - 2\,\varphi_{ij}}{4 - \varphi_{ii} - \varphi_{jj} - 2\,\varphi_{ij}}. \tag{4}$$

No phylogeny is implied by a dendrogram, although if the number of loci on which it is based is sufficiently large, the major branches (and less reliably, the minor branches) may have phylogenetic significance.

A phyletic tree, or "cladogram," is an interpretation of a dendrogram in phylogenetic terms, and its meaningful axis represents time in years or generations. The time scale may be inferred from paleontological evidence or, less accurately, by a transformation of hybridity, assuming a uniform divergence rate; a more realistic model founders on our ignorance of evolutionary sizes and systematic pressures

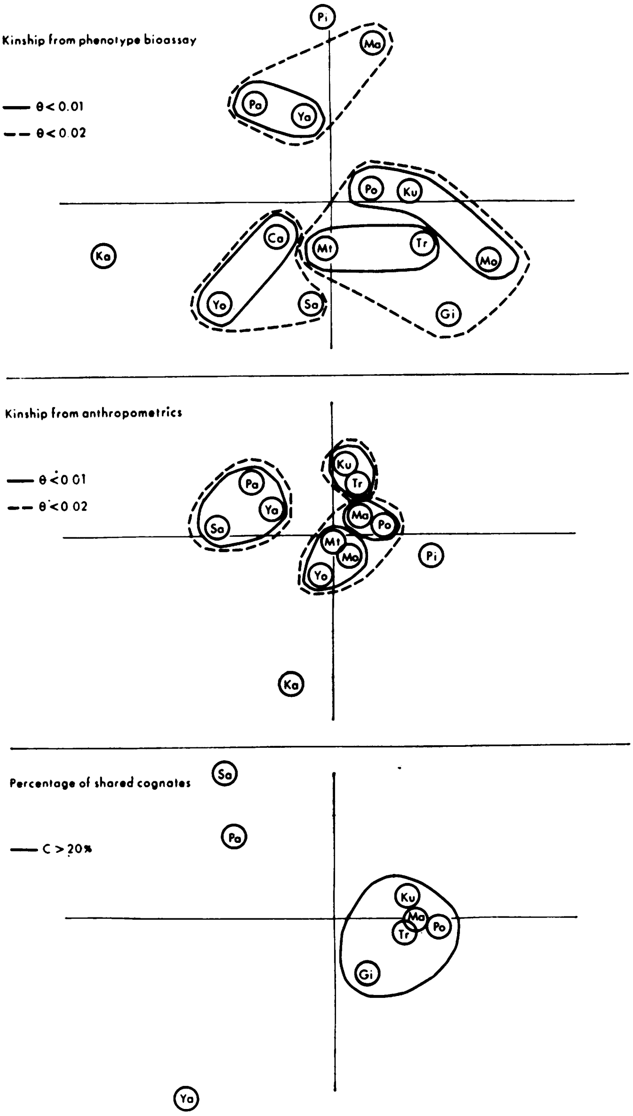

FIG. 2.—Eigenvectorial representations rotated to maximum congruence with geography

during differentiation. We suppose initially with glottochronology that contemporary kinship is

$$\varphi_{ij} = ae^{-B(t^* - t_{ij})}, \qquad (5)$$

where a is the Malécot intercept for the first branch, t^* is the duration of the array as the time since the first split, t_{ij} is the time from the origin of the array to differentiation of populations i and j, and $a,B > 0$. Assume a and t^* known. Then t^* is associated with a minimum kinship φ^* and a maximum hybridity θ^*. If we set $\varphi_{ii} = \varphi_{jj} = a$ in equation (4) and let φ_{ij} approach zero, we see that $\theta^* \doteq a/2$. Substituting $a = 2\,\theta^*$ and rearranging,

$$\varphi_{ij} = \frac{2\,\theta^*(1+\theta_{ij}) - 2\,\theta_{ij}}{1-\theta_{ij}}, \tag{6}$$

and the time from the present back to differentiation of populations i and j is

$$t^* - t_{ij} = t^* \left\{ \frac{\ln(2\,\theta^*/\varphi_{ij})}{\ln[(1-\theta^*)/\theta^*]} \right\} \text{ b.p.,} \tag{7}$$

where "b.p." signifies "before present."

The roughly exponential decline of cognate frequencies with time may be an artifact of an increasing proportion of cognates with high retention rate, combined with an accelerating pressure toward diversification. The latter also seems likely for genes, since increasing time implies increasing distance and greater exposure to drift and disruptive selection. As an approximation, we might assume that under diversifying pressure equation (5) becomes [17]:

$$\varphi_{ij} = ae^{-B(t^*-t_{ij})^2}, \tag{8}$$

in which case equation (7) is replaced by

$$t^* - t_{ij} = t^* \sqrt{\frac{\ln(2\,\theta^*/\varphi_{ij})}{\ln\,[(1-\theta^*)/\theta^*]}} \text{ b.p.} \tag{9}$$

The value of θ^* is determined in constructing the dendrogram. One method seeks to maximize θ^* by considering all nested partitions of the n populations, which is enormously expensive of computer time. This kind of cluster analysis was developed to minimize the cost of communication systems; its relevance to evolution is unclear, although it is often combined with the pretension that the dendrogram is in every detail a cladogram of populations evolving at a constant rate [9]. Without this assumption, the method has not been shown to have any desirable properties. An alternative procedure [10, 11] clusters populations with the smallest value of hybridity and thus proceeds from the smallest to the largest branches without testing all nested partitions and without assuming that the dendrogram is a cladogram. Since trees permit branching but not anastomosis, and every human population undergoes "fusion" (i.e., hybridization), it cannot be argued that a dendrogram of subspecific differentiation is in every branch a cladogram, even if the number of bioassays were sufficiently large that sampling error of θ could be neglected. It is worth noting that the arbitrary but reasonable standard of numerical taxonomists that at least 100 characters be assayed is never approached by geneticists, and appreciable sampling errors must be assumed and have in fact been simulated [12]. Therefore, we prefer the phenetic approach, which makes fewer assumptions, requires fewer calculations, and is almost universally employed in numerical taxonomy, which has always insisted that its trees are not cladograms

[10]. Not all geneticists have appreciated this sophistication. Parenthetically, it is improper to call a tree a network, since the latter permits anastomosis as well as branching. It is precisely this restriction on trees, together with the effect of sampling errors on tree form, that makes the interpretation of a dendrogram as a cladogram generally unconvincing. However, even when the tree form is incorrect, a cladogram yields estimates of divergence times of local populations that may sometimes be interesting.

Having rejected the phylogenetic assumption as a principle for constructing trees of subspecific differentiation, we have one further decision to make: how are populations to be weighted in forming averages? The greatest statistical reliability is obtained by weighting each population independently of the way in which populations are grouped into branches. These a priori population weights contrast with a posteriori stem weights, in which every branch is weighted equally, independently of the number of populations that comprise it [10]. We can see no justification for stem weights in constructing a dendrogram.

The algorithm used by us for our computer program ARBOR is the so-called unweighted pair-group method [10]. A dendrogram is constructed by successively reducing the rank of the hybridity matrix, at each stage averaging the rows and columns corresponding to the smallest value of θ, and deleting the row and column with the higher index. The weights are simultaneously summed. At the end, when the hybridity matrix contains only θ^*, a dendrogram and (if t^* was specified) the derived cladogram are graphed.

The formula analogous to equation (7) for cognates is

$$t^* - t_{ij} = t^* \left[\frac{\ln C_{ij}}{\ln C^*} \right] \text{ b.p.,} \qquad (10)$$

where C_{ij} is the cognate frequency between populations i and j and C^* is the cognate frequency for populations that diverge at t^*. If we take $\theta = 1 - C$, then C^* is the complement of θ^*, the last value left in the reduced matrix. Glottochronology traditionally assumes that $t^*/\ln C^* = -2{,}300$, based on large continental populations, but the generality of this constant is doubtful [13].

The cognate formula analogous to equation (9) is

$$t^* - t_{ij} = t^* \sqrt{\frac{\ln C_{ij}}{\ln C^*}} \text{ b.p.} \qquad (11)$$

In place of cognates we may use similarity, which codes a consistent-trait pair (++ or −−) as one and an inconsistent pair (−+ or +−) as zero. The complement, dissimilarity, may be treated as hybridity in constructing a dendrogram. Thus, the same algorithm can be applied to cultural traits, cognates, phenotypes, and anthropometrics.

Dendrograms with equal weights are given in figure 3. Both phenotypes and anthropometrics indicate a primary division between the Polynesian outlier, Kapingamarangi, and the Micronesian populations, within which Pingelap is most

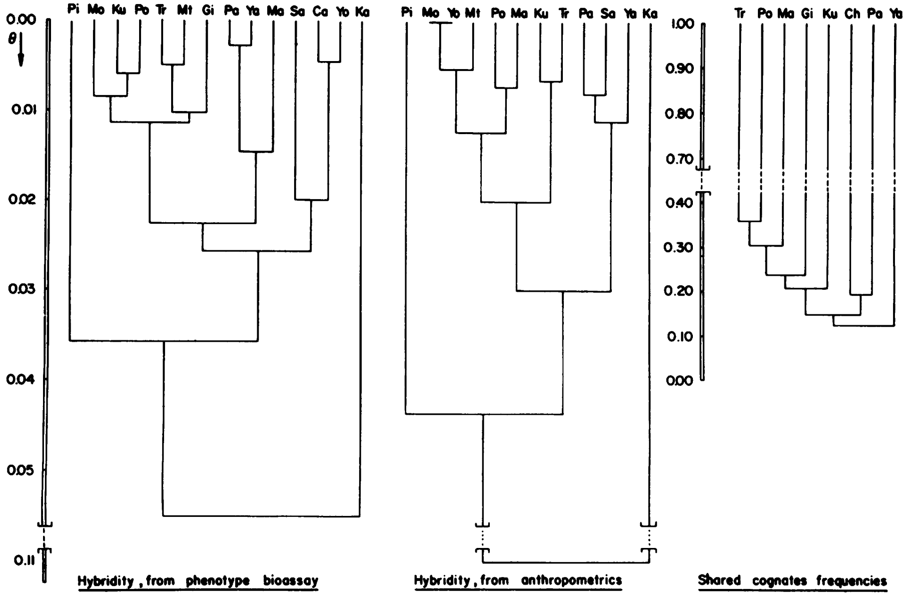

FIG. 3.—Dendrograms of hybridity in Micronesia

divergent. Other Eastern Caroline populations form a rather homogeneous group. There is an appreciable east-west divergence within Micronesia, in agreement with the cognate matrix.

Recent archeological evidence from Tonga gives radiocarbon dates of 3,000 years, which we take as t^* on the assumption that divergence of Polynesians and Micronesians occurred at about that time [14]. Equation (7) is not compatible with this assumption, which gives ridiculously small estimates of divergence times. For example, phenotypes yield 50 years for the separation of Yap and Palau. Quadratic differentiation [eq. (9)] leads to much more plausible results (fig. 4). The separation time for Yap and Palau based on phenotypes becomes 386 years. Divergence of Pingelap from the rest of eastern Micronesia is estimated to have occurred 1,711 years ago from phenotypes and 1,323 years ago from anthropometrics. However, divergence time for Mokil is given as only 692 years from phenotypes and 398 years from anthropometrics. It is unlikely that Mokil was occupied less than half as long as Pingelap. In a cladogram a population that is unusually differentiated because of small evolutionary size or systematic pressure tends to be assigned a spuriously high estimate of divergence time. If it is assumed that Pingelap and Mokil were settled at about the same time, we take the mean of these four numbers, which is 1,031 years, in reasonable agreement with estimates of occupancy from kinship [1].

The results for cognates are not strictly comparable since data on Kapingamarangi and Pingelap were not included. If we retain the traditional assumption of linear divergence [eq. (10)] and $t^* = 1{,}500$, the separation of Trukese from

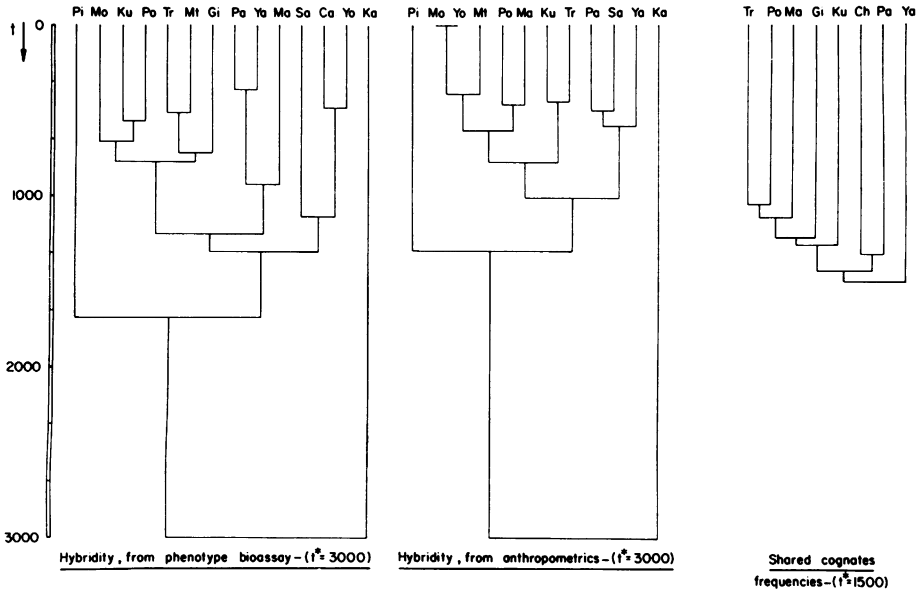
t
0
1000
2000
3000
Pi Mo Ku Po Tr Mt Gi Pa Ya Ma Sa Ca Yo Ka
Pi Mo Yo Mt Po Ma Ku Tr Pa Sa Ya Ka
Tr Po Ma Gi Ku Ch Pa Ya
Hybridity, from phenotype bioassay – (t^*= 3000)
Hybridity, from anthropometrics – (t^*= 3000)
Shared cognates frequencies – (t^*=1500)

Fig. 4.—Cladograms of hybridity in Micronesia

Ponapean is dated as 734 b.p. compared with 1,049 b.p. for quadratic differentiation, which gives 811 b.p. from phenotypes and 810 b.p. from anthropometrics. Although linear differentiation is usually assumed for cognates, quadratic differentiation is at least as plausible, since migration rates are likely to decrease with distance, and therefore time. More experience is needed with dating from cladogram models before any validity may be claimed for them.

DISCUSSION

One of the problems in bioassay of kinship is whether to express the results as φ_{ij} with respect to the contemporary array (random kinship of zero) or as φ_{ij}' relative to founders (positive random kinship), the two systems being related as $\varphi_{ij}' = (\varphi_{ij} - L)/(1 - L)$, where L is the initial estimate at large distances [1]. Malécot [15] calls φ_{ij} the "conditional kinship" and φ_{ij}' the "a priori kinship." We have used φ_{ij} as input for topology, so that our results do not depend on estimation of L.

In this region of long-range voyages [16], it would be naïve to consider that the dendrogram is a cladogram. Predictions of kinship from a migration matrix can be represented by a dendrogram, but the assumptions are incompatible with a cladogram. Migration theory assumes that populations have evolved under hybridization from synchronous random samples of an infinite gene pool, while a cladogram assumes random subdivision of a finite gene pool at different times without hybridization: kinship increases monotonically under the first model and decreases after fission under the second. These essentially different processes were not dis-
"

tinguished by Cavalli-Sforza in his discussion of the evolution of kinship [17]. Hybridization, accidents of drift, and diversifying selection are powerful obstacles to interpreting a dendrogram as a cladogram. The kinship matrix itself has errors, and some of the information it contains must be lost in any graphical representation.

With this caveat, we can compare topologies on the basis of the cophenetic correlation between observation and prediction [18]. This correlation has the merit of being invariant with respect to scalar addition or multiplication and so is the same for a priori and conditional kinship. For principal eigenvectors the predicted kinship is $E_1P_1 + E_2P_2$, where E_1, E_2 are the principal eigenvalues and P_1, P_2 are the outer squares of the vectors V_1, V_2, respectively. We have correlated hybridity rather than kinship so as to compare with dendrology, which predicts hybridity as the mean of a submatrix.

From table 1 it is apparent that principal eigenvectors have the higher cophenetic correlation for phenotypes, and in this sense the dendrogram is a poorer two-dimensional graph of kinship. It has been found that population weights tend to give higher cophenetic correlations than stem weights [19] and that dendrograms constructed by the minimum variance principle give lower cophenetic correlations and suffer from other theoretical disadvantages [18–21] in addition to their heavier computational load, which is inevitable in any minimization procedure on trees. We are not convinced that this extra labor has either theoretical or practical justification.

TABLE 1

RELIABILITY OF TOPOLOGIES

	PRINCIPAL EIGENVECTORS		DENDROLOGY
SOURCE	$f(\lambda)$	r	r
Phenotypes	.621	.927	.746
Metrics	.685	.906	.944
Cognates	.341	.834	.948

NOTE.—r = cophenetic correlation; $f(\lambda)$ = proportion of absolute dispersion accounted for by two principal eigenvectors.

Taxonomic applications favor dendrograms in which each split divides populations subequally, as opposed to chained dendrograms, which tend at each stage to separate the most divergent population from the remainder. For genetics the goal is to represent the kinship matrix in two dimensions, and so there is no objection to chaining.

In Micronesia eigenvectorial representations and dendrograms both provide a simple representation of kinship. We suggest that both representations be examined when summarizing the information in a kinship matrix. If the investigator recognizes these topologies as a convenient distortion, he will be less likely to indulge in unwarranted phylogenetic speculation. However, since neither representation is isometric, the relations among populations are necessarily distorted. Only the original kinship matrix contains all the information about differentiation.

SUMMARY

Bioassay of Micronesian kinship from phenotypes, anthropometrics, and cognates has been represented in two dimensions by principal eigenvectors and dendrograms. These methods sacrifice some of the information in a kinship matrix to achieve a simple graph but conserve a large part of the information. Together they show the Polynesian outlier Kapingamarangi to be highly differentiated from the Micronesian populations, among which Pingelap is most divergent, as predicted from migration and genealogy. There is marked east-west differentiation, but other details are not consistent in the different representations. It is recommended that both topologies be examined when analyzing a kinship matrix, to avoid unwarranted phylogenetic speculation. An algorithm is introduced for transforming a dendrogram into a cladogram, which gives fairly plausible estimates of divergence times. Difficulties in interpreting a dendrogram as a cladogram are discussed.

REFERENCES

1. MORTON NE, LALOUEL JM: Bioassay of kinship in Micronesia. *Amer J Phys Anthrop.* In press, 1973
2. IMAIZUMI Y, MORTON NE: Isolation by distance in New Guinea and Micronesia. *Archaeology Phys Anthrop Oceania* 5:218–235, 1970
3. MORTON NE, MIKI C, YEE S: Bioassay of population structure under isolation by distance. *Amer J Hum Genet* 20:411–419, 1968
4. HAINLINE LJ: Population and genetic (serological) variability in Micronesia. *Ann NY Acad Sci* 134:639–654, 1966
5. POLLOCK N, LALOUEL JM, MORTON NE: Kinship and inbreeding on Namu Atoll. *Hum Biol* 44:459–474, 1972
6. LALOUEL JM: Topology of population structure. In preparation
7. SCHÖNEMANN PH, CARROLL RM: Fitting one matrix to another under choice of a central dilation and a rigid motion. *Psychometrika* 35:245–255, 1970
8. MORTON NE, YEE S, HARRIS DE, LEW R: Bioassay of kinship. *Theor Pop Biol* 2:507–524, 1971
9. CAVALLI-SFORZA LL, EDWARDS AWF: Phylogenetic analysis: models and estimation procedures. *Amer J Hum Genet* 19:233–257, 1967
10. SOKAL RR, SNEATH PHA: *Principles of Numerical Taxonomy.* San Francisco, W. H. Freeman, 1963
11. HARPENDING H, JENKINS T: Genetic distance among South African populations. In preparation
12. KIDD K, CAVALLI-SFORZA LL: Error in the reconstruction of evolutionary trees, in *Genetic Distance*, edited by CROW JF, New York, Plenum. In press, 1972
13. HYMES DH: Lexicostatistics so far. *Current Anthrop* 1:3–44, 1960
14. GROUBE LM: Tonga, Lapita pottery, and Polynesian origins. *J Polyn Soc* 80:278–316, 1971
15. MALÉCOT G: Structure géographique et variabilité d'une grande population, in *Human Genetics,* Proceedings 4th International Congress of Human Genetics, Paris, September 1971, Amsterdam, Excerpta Medica, 1972, pp 138–154
16. RIESENBERG SH: Table of voyages affecting Micronesian Islands. *Oceania* 36:155–170, 1965

17. CAVALLI-SFORZA LL: Human diversity, in *Proceedings 12th International Congress of Genetics*, Tokyo, August 1968, vol 3, 1969, pp 405–416
18. SOKAL RR, ROHLF FJ: The comparison of dendrograms by objective methods. *Taxon* 11:33–40, 1962
19. SNEATH PHA: Evaluation of clustering methods, in *Numerical Taxonomy*, edited by COLE AJ, New York, Academic Press, 1969, pp 257–271
20. WISHART D: Mode analysis: a generalization of nearest neighbor which reduces chaining effects, in *Numerical Taxonomy*, edited by COLE AJ, New York, Academic Press, 1969, pp 282–311
21. BOYCE AJ: Mapping diversity: a comparative study of some numerical methods, in *Numerical Taxonomy*, edited by COLE AJ, New York, Academic Press, 1969, pp 1–31