Report on Nineteenth International Numerical Taxonomy Conference

Pierre Legendre

Systematic Zoology, Vol. 35, No. 1 (Mar., 1986), 135-139.

Points of View

Report on Nineteenth International Numerical Taxonomy Conference

PIERRE LEGENDRE

Département de Sciences biologiques, Université de Montréal,
C.P. 6218, Succursale A, Montréal, Québec H3C 3J7, Canada

The Nineteenth International Numerical Taxonomy Conference (NT-19) was held at Université de Montréal, 4–6 October 1985. The host of the meeting was David Sankoff, from the Centre de Recherches mathématiques. Fourteen papers were presented, and the number of registered participants was 50. Since the meeting was not divided into theme-oriented sessions, the following report will shuffle the talks to present them under three major headings: Methods (with half of the presentations), Computers, and Applications. Certainly, the most significant methodological development was the spreading of randomization tests (used in four of the papers) as an approach for solving a variety of NT problems.

1. Methods.—The first paper of the meeting was by G. D. Schnell, J. N. Thompson, and J. J. Hellack (Univ. Oklahoma, Norman) and entitled "Numerical Techniques for the Analysis of Polygenes." Its methodological contribution resided in the use of the Kolmogorov-Smirnov Z-statistic to measure the distance between frequency distributions of quantitative characters, representing the phenotypic expression of polygenic factors in different isofemale strains. While the authors' aim was clearly to quantitatively compare these frequency distributions to one another before clustering and ordination, discussion from the floor centered on whether it was appropriate to lump together all sources of variation (central tendency, dispersion, and skewness), rather than to analyze them separately. As an overall measure, the Z-statistic performed well; its use could be combined with other techniques that decompose variation into separate components.

In "Multiple Regression as a Cladistic Tool: Plesiomorphy and the Common Factor," J. McNeill and K. Conlan (Univ. Ottawa and Carleton Univ., Ottawa) applied in the cladistic context Wood's (1983) phenetic method for removing the *common part* of a set of OTUs before carrying out the tree reconstruction. McNeill argued that this common factor may be interpreted in cladistic terms as the plesiomorphic component of the clade. In this method, one regresses in turn, in Q-mode, each OTU of the ingroup against the members of the outgroup, and computes the residuals. Only the residuals are then used for phylogenetic analysis, which was carried out in this case by computing a Steiner tree. This method was suggested as a possible tool for handling continuous variables in cladistic analysis; it was applied to 23 continuous characters from Conlan's (1986) study of the amphipod genus *Microjassa*, where it gave a closer fit to the cladogram obtained from the complete data set of 104 characters than did a cladistic analysis using these same 23 continuous characters converted to discrete states by gap coding

followed by additive binary coding. McNeill's presentation led to a healthy discussion on the validity of the method of Q-mode regression, and especially on the invariance of residuals under changes in the clade's membership (if the characters are first standardized) or under arbitrary changes in measurement scales (if they are not standardized).

R. R. Sokal (State Univ. New York, Stony Brook) presented advances in the use of Mantel tests in a paper entitled "Normalized, Categorical, and Multiple Mantel Tests." The Mantel test is appropriate for testing the correlation among variables when the samples are not independent of one another because the phenomenon under study displays spatial or temporal autocorrelation. Using two distance matrices X and Y to represent the relationships between all object pairs for the two variables to be tested, Mantel's Z-statistic is simply the sum of cross-products of the terms in the two matrices. Y may represent the geographic distances among objects. The significance of Z is tested against a reference distribution obtained by repeated permutations of the object order in one of the two distance matrices. Sokal first addressed the problem of obtaining a spatial correlogram using Mantel's statistic, instead of the usual Moran's I or Geary's c statistics. This can be done by making one of the two distance matrices, say Y, contain ones only for pairs of objects located within the distance class being analyzed, and zeroes elsewhere. A Z-statistic is computed in turn for each distance class. Before a correlogram can be drawn out of these results, the various Z-statistics have to be brought down to a common scale, which can be done in one of two recently developed ways: by transforming Z into a correlation coefficient (Smouse, Long, and Sokal, in prep.), or by scaling it between 0 and 1 using the expectation of Z, its minimum, and its maximum values (Hubert et al., 1981). Since the Mantel test can handle distance matrices based on many variables, as well as on single variables, this represents an important advance to methodology, because it makes it possible to

study directly the spatial structure of multivariate data sets such as genetic distances; besides taxonomy, this method should also prove extremely useful in population genetics and in ecology, among other fields, where the relevant information is essentially multivariate. The second problem discussed by Sokal was how to compute a correlogram from categorical data. The solution proposed consists of writing ones in distance matrix X for those joins of interest, such as paired like or unlike states, while ones in matrix Y represent pairs of objects located in a given distance class, as before. Finally, Sokal showed how to compute partial correlations between two distance matrices A and B, holding a third one (C) constant (Dow and Cheverud, 1985; Hubert, 1985; Smouse, Long, and Sokal, in prep.)

In a paper entitled "Clustering Without Similarities: Lots of Headaches," L. P. Lefkovitch (Agriculture Canada, Ottawa) presented a clustering method based on presence/absence data that does not require the previous computation of a resemblance matrix, as is usually the case in clustering. The method consists of phases 2 and 3 of *conditional clustering* that Lefkovitch presented to this group at NT-17; rather than starting from a pairwise resemblance matrix as in conditional clustering, the maximum entropy principle is used to obtain the probabilities of the various possible groups directly from the binary raw data. These probabilities make it possible to compute an optimal set covering (which can be a partition), as in conditional clustering. Lefkovitch described a number of headaches generated by this method, which is still under development, and showed applications to ecology and psychology.

G. F. Estabrook (Univ. Michigan, Ann Arbor) expanded on "Convex Phenetics," following the concept of *convexity* that he had introduced during the NATO Advanced Study Institute on Numerical Taxonomy in 1982 (Estabrook, 1983). Convexity is achieved if, for every pair of objects that are members of a subset, every OTU that is located *between* these two objects is

also a member of the subset. In phylogeny, *between* refers to any HTU located along the path of ancestors and descendants that connects two OTUs or HTUs. Estabrook proposed to constrain the building of the phenetic tree, by using the information available from a phylogenetic tree. The constraint actually consists of disallowing partitions whose classes cannot all be convex on the phylogenetic tree. Convex phenetics has many advantages: it reduces the instability of cladistic classifications; it recognizes phenetic similarity and divergence, and in this sense it leads to a *phylogenetic* classification (sensu Mayr, 1985); it obeys the principle of convexity, which is philosophically appealing. Its results may differ depending on the clustering algorithm being used, but less so than in purely phenetic studies, because many inappropriate solutions are eliminated by the phylogenetic convexity constraint.

In a paper entitled "Relationships Among Quartet Dissimilarity Measures Between Undirected Phylogenetic Trees," W. H. E. Day (Memorial Univ. of Newfoundland, St. John's) presented six measures of dissimilarity between trees, based upon quartet counts. These included the four measures presented by G. F. Estabrook in 1984 during the NT-18 Conference (Estabrook et al., 1985). A quartet is a subtree of four OTUs; the various measures of dissimilarity between trees are all based on the number of like and unlike quartets in two trees of the same OTUs. By Monte Carlo generation of trees, Day investigated the statistical properties (mean and standard deviation) of the six measures; this information will allow users of these measures to decide whether a given value, obtained for a pair of real trees, is large or small.

J. W. Archie (Univ. Hawaii, Honolulu) presented "A Randomization Test for the Presence of Cladistic Structure, using the Consistency Index." He first described the behavior of the Kluge and Farris (1969) consistency index, C, as the number of taxa or the number of characters increases, and as the shape of the tree changes in data sets with randomly assigned character states. He finally suggested a randomization test permitting one to decide whether the value of C obtained in any particular case is sufficiently large to allow rejecting the null hypothesis of it being obtained from data with randomly assigned states. The reference distribution of C is obtained by permuting the character-state assignments among OTUs, computing a minimum-length tree, and calculating C; this process is repeated a number of times. Archie pointed out that this method is limited by the fact that it is very easy to reject H_0, because essentially all minimum length trees obtained from real data are expected to display more structure than trees from randomly assigned states.

2. Computers. —J. Felsenstein (Univ. Washington, Seattle) presented a paper on "The PHYLIP Phylogenetic Inference Package: Its Past, Present, and Future." This package, known to most of us who use numerical methods to infer phylogenies, has been distributed, free of charge, to over 350 installations around the world, since its first release in October 1980. It consists at present of 18 programs covering all the general-purpose methods for inferring phylogenies, plus various documentation files. It is written in a dialect that Felsenstein described as *Paranoid Pascal*, that is, a subset of Pascal where everything has been taken out that might cause trouble on any one particular machine. The present release, version 2.7, includes several new features that should make anyone want to get the latest release.

Two talks centered on algorithms to solve the problem of exhaustively searching for the most-parsimonious tree of a set of OTUs. In their talk entitled "Parallel Processing and Parsimonious Phylogenies," D. Sankoff, Y. Abel and R. J. Cedergren (Univ. Montréal) worked on developing an algorithm using parallel processing, on a CRAY 1-S computer. Parallel processing seems to be the most promising method for saving time, when examining all trees. Not all the potential saving of time was achieved, however, because of the awkward programming imposed by parallel processing. The present

algorithm can handle up to 9 OTUs with 100 five-value characters, in 13 min of CPU time.

Mariana Constantinescu (Univ. Moncton) used another approach in her presentation entitled "Enumerating Trees Under Constraints." She reduced the number of trees to be searched by imposing known structural constraints on the set of trees to be scanned. These constraints may often be formulated in terms of a consensus tree or forest. She stated a theorem which gives a reduced number of trees as a function only of the degrees of the vertices in the consensus trees.

3. *Applications.*—Presenting a paper prepared by J. Moret, J.-M. Hubac, H. Couderc and R. Gorenflot (Univ. Paris XI, Orsay), Hubac described some "Applications of Numerical Analysis to the Identification of Taxa in Genus *Ornithogalum* L. (Subgenus *Beryllis*)." The problem was to determine whether one could identify these species (from Morocco) using morphological data alone, despite their large phenotypic variability. This was achieved using principal component analysis and hierarchical clustering; the resulting clusters were corroborated by ecological as well as chromosome number data.

R. J. Jensen (Saint Mary's College, Notre Dame, Indiana) presented a talk on "Phenetic and Geographic Spatial Autocorrelation in Oaks: An Evaluation of Fruit Characters." Jensen used eight size measures of the acorns, plus nine ratios of these size characters used as descriptors of shape, for a total of 17 characters. He first computed Estabrook and Gates' (1984) *goodness* coefficient (called *phenetic spatial autocorrelation* by Estabrook and Gates). It was pointed out during the discussion that Estabrook and Gates' *goodness* is nothing but a Mantel statistic (see above) between a distance matrix for a single character, elevated to an arbitrary power Y, and a Minkowski distance matrix among OTUs based on all the remaining characters, to the power X; one studies the goodness of characters for different values of powers X and Y. Jensen noticed that a character's goodness rank not only varied as a func-

tion of X and Y, but also depended on the method used to standardize characters. In the second part of his paper, Jensen computed spatial autocorrelation using Moran's I coefficient, based on a Gabriel graph connecting the sampling stations. He found that his size measurements were significantly autocorrelated in space, but not his shape characters. Discussion centered on the use of ratio characters in such studies, and on the inappropriateness of using the number of edges of a Gabriel graph as the measure of distance in autocorrelation analysis, instead of geographic distance, when the position of the points in space is determined solely by the worker and does not reflect particularities of the distribution of the objects under investigation.

W. S. Modi (Laboratory of Viral Carcinogenesis, National Cancer Institute, Frederick) presented "Phylogenetic Analyses of Chromosomal Banding Patterns Among Nearctic Arvicolidae (Mammalia: Rodentia)." After obtaining metaphase karyotypes for 22 species of rodents (2n = 18 to 64) with high-quality G-banding, Modi found that 45 homologous chromosomes or chromosomal arms were shared by two or more species, and that 139 euchromatic rearrangements accounted for the extensive karyotypic variability. The methodological contribution of this paper lies in Modi's development of a clever way to code chromosomal rearrangements into binary characters and weigh them variably relative to entire chromosomes (Modi, 1986). The resulting data matrix (22 species × 221 chromosomal characters) was then subjected to a Wagner parsimony analysis in order to get a cladistic tree. The results were discussed and compared with classifications derived from paleontology, comparative morphology, electrophoresis, and zoogeographic distributions.

S. E. Hartman (State Univ. New York, Stony Brook) presented "A Molar Perspective on Hominoid Systematics." First, he noted that because of their prevalence in the fossil record, teeth are often used in discussions on the evolutionary systematics of vertebrates. Wondering whether

teeth may be used alone for phylogenetic inference, Hartman assembled a data set consisting of the X, Y and Z coordinates of 34 molar landmarks, for 250 specimens representing six extant hominoid taxa. The size-adjusted data were examined phenetically using canonical variates analysis, and lengths for all 105 possible trees were obtained by Wagner parsimony. It appeared that diet, rather than evolutionary propinquity, was the chief determinant of dental affinities found by both methods. This result raised interesting questions as to the reliability of teeth as phylogenetic indicators.

The future of numerical taxonomy was discussed by Joseph Felsenstein during his banquet address, entitled "Contemporary Systematics: A View From Outside." He convincingly argued that the major problems now faced by biological systematics would find a solution in the next quarter of a century, and that this solution would be based upon the vast number of nucleic acid sequences that are presently being decrypted by molecular evolutionary biologists.

The NT-20 Conference will be held in October 1986 at the State University of New York at Stony Brook. R. R. Sokal and F. J. Rohlf (Department of Ecology and Evolution), along with N. Creel (Department of Anatomical Sciences), will host the meeting. It was suggested that this conference could offer symposia, for discussion of subjects of general interest to taxonomists, in addition to the contributed paper sessions. Details will be announced in a forthcoming issue of *Systematic Zoology*. NT-21 is scheduled to be held at the Zoological Station of the University of Washington, in Friday Harbor, Washing-

ton, 23 to 25 October 1987. Plans have also been made to hold NT-22 at Berea College, Berea, Kentucky in October 1988.

REFERENCES

CONLAN, K. 1986. Phenetic and cladistic methods applied to a small genus (Corophioideae: Ischyroceridae: *Microjassa*) and a larger outgroup. Crustaceana (in press).

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

ESTABROOK, G. F. 1983. The causes of character incompatibility. Pages 270–295 *in* Numerical taxonomy. Proceedings of a NATO Advanced Study Institute (J. Felsenstein, ed.). NATO Advanced Study Institute Series G (Ecological Sciences), No. 1. Springer-Verlag, Berlin.

ESTABROOK, G. F., F. R. MCMORRIS, AND C. S. MEACHAM. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Syst. Zool., 34:193–200.

ESTABROOK, G. F., AND B. GATES. 1984. Character analysis in the *Banisteriopsis campestris* complex (Malpighiaceae), using spatial auto-correlation. Taxon, 33:13–25.

HUBERT, L. J. 1985. Combinatorial data analysis: Association and partial association. Psychometrika, 50:449–467.

HUBERT, L. J., R. G. GOLLEDGE, AND C. M. CONSTANZO. 1981. Generalized procedures for evaluating spatial autocorrelation. Geograph. Anal., 13:224–233.

KLUGE, A., AND J. S. FARRIS. 1969. Quantitative phyletics and the evolution of anurans. Syst. Zool., 18:1–32.

MAYR, E. 1985. Darwin and the definition of phylogeny. Syst. Zool., 34:97–98.

MODI, W. S. 1986. Phylogenetic analyses of chromosomal banding patterns among the Nearctic Arvicolidae (Mammalia: Rodentia). Syst. Zool., 35 (in press).

WOOD, D. S. 1983. Character transformations in phenetic studies using continuous morphometric variables. Syst. Zool., 32:125–131.