SYSTEMATICS ASSOCIATION PUBLICATION NUMBER 6

PHENETIC AND PHYLOGENETIC CLASSIFICATION, pp. 67-76

Issued DECEMBER 1964

Reconstruction of Evolutionary Trees

A. W. F. EDWARDS & L. L. CAVALLI-SFORZA

(*International Laboratory of Genetics and Biophysics, Naples (Pavia Section), Istituto di Genetica, Università di Pavia, Italy*)

CONTENTS

1. INTRODUCTION

THE extent to which the classification of a group of organisms may be used to make inferences about phylogenetic relationships has long been a controversial point in biology. Today, particularly in view of the development of 'numerical taxonomy' it is important to consider, more deeply than hitherto, the whole question of evolutionary inferences and their relation to taxonomy. This paper contains some suggestions for studying the problem, which lead to practical methods of estimation. First, however, it is convenient to consider the logical nature of taxonomic studies in general.

2. TAXONOMY

We suspect that the present confusion of ideas about the logical basis of taxonomy is partly due to a failure to distinguish between different types of taxonomic enquiry, and that this failure may have been induced by the fact that all taxonomic studies use much the same type of data. Every kind of taxonomic enquiry must have its own particular purpose and logical basis; indeed, the logical structure must arise naturally from the particular purpose, which therefore needs to be specified in detail. A classification created with no particular purpose in mind is likely to be logically indefensible and partially useless; for when it is put to a specific* use, for which it was not, of course, designed, it will

* We use *specific* and its derivatives in their everyday senses.

most probably be inefficient. In the same way, a motor vehicle built without a specification is unlikely to turn out to be a fast sports car: it is more likely to be a heterogeneous collection of ill-suited mechanical bits and pieces.

Though the need to specify the purpose of a taxonomic study has, from time to time, been admitted, some confusion has been generated by the concept of the 'general purpose' classification. To say that the purpose of a classification is "general" is, in our view, too vague to be of use in its construction. If we may return to the motor vehicle, a 'general purpose' car can be specified only because the particular purposes to which it may be put, and their relative importance, can themselves be specified rather accurately. Thus the designer may anticipate that the proposed vehicle will spend 50% of its time carrying only the driver, 20% carrying two people, and so forth, and that it will be required to carry so much in the way of goods so often, and go up hills of such-and-such a gradient once a year on average. On the basis of such information an acceptable compromise design will be achieved. There are three important things to note in this example. First, that 'general purpose' is only meaningful if it is regarded as a collection of weighted specific purposes: secondly, that the outcome of a general purpose specification is unlikely to fulfil any of the specific purposes as well as a specially designed vehicle could; and, thirdly, that the justification of the general purpose vehicle is an economic one, because for optimum service all the time one would need as many vehicles as there are uses, but this would be very expensive: only rich farmers can afford a Jaguar *and* a Land-Rover.

In taxonomy the situation is much the same; the only justification for a general purpose classification is that we cannot afford the time or the money to create the separate specific classifications that we really need. It seems to us to have no purely scientific justification, and it is surely an illusion to imagine that a 'general purpose' classification approaches an unattainable 'ideal' one in the same way that a circle drawn on a piece of paper approaches a perfect circle. We thus hold that the most essential prerequisite of a classification is its purpose, clearly defined, and that this is no less true in the case of a 'general purpose' classification, which should only seek to be a compromise solution to the problem of satisfying specific purposes economically.

Given the purpose to which a classification is to be put, its design, the choice and weighting of the characters to be used, and indeed its logical structure, follow, just as the design of a motor car follows from its specification. Naturally, there is likely to be room for subjective judgements in the choice and weighting of characters, and we must not be surprised to find two taxonomists disagreeing even when the specification of the classification seems quite rigid, just as there is room for motor car designers to interpret a specification subjectively. But the scope for subjective decisions is directly related to the comprehensiveness of the specification, which should be complete if objectivity is very important. Conversely, when there is virtually no specification

at all, as in the 'general purpose' case as used hitherto, there is immense scope for subjective decisions, and attempts to find a logical justification where none exists are doomed to death by drowning in a sea of indefensible statements.

Rather as Fowler apologizes for taking all his examples of bad English from *The Times*, and insists that this is not because *The Times* contains more bad English than most newspapers, but because it is the newspaper he prefers, so we apologize for taking some statements about general purpose classification from the admirable summary of Numerical Taxonomy by Sneath and Sokal (1962) and quoting them:

"Such [natural] classifications may be termed 'general' and are for general purposes, in distinction to classifications made with a special purpose in mind, which may be called 'special' classifications. This implies among other things that in constructing general classifications characters have equal weight...".

"The taxonomic equivalence of all characters can be seen most clearly if we attempt to construct an objective criterion for weighting characters. If we cannot decide how to weight the features we must give them equal weight, unless we propose to allocate weight on irrational grounds".

"We hold the view that a 'natural' or orthodox taxonomy is a general arrangement intended for general use by all kinds of scientists; it cannot therefore give greater weight to features of one sort, or it ceases to be a general arrangement".

The logical frailty of these attempts to justify a certain procedure when the purpose of the intended classification has not been carefully specified should be a warning to us all.

There seem to us to be three major types of taxonomic study, each of entirely different logical structure, and it is of the utmost importance to differentiate them. The first, which we may call Pure Taxonomy, is solely concerned with the processing of information. It is descriptive. Each organism can only be described in terms of the characters it exhibits, but since some of these characters will be common to two or more of the organisms, it is clearly unnecessary to describe each organism separately, and some form of classification will present all the information in a simpler form. This approach has been developed by Maccacaro (1958) and Rescigno and Maccacaro (1961), and has not, in our view, received the attention it deserves. It should be noted that the problem of choosing and weighting characters is irrelevant in *pure* taxonomy: *given* the characters, the problem is to classify the organisms according to them in such a way that the maximum amount of information is retained. Thus, given a set of snooker balls, it is possible to classify the balls by colour without loss of any information about this character.

The second type of taxonomy may be called "applied", for it seeks to classify the organisms in such a way that the resulting classification shall be the best obtainable for the specified purpose at hand. Thus tyre manufacturers are primarily interested in classifying motor vehicles by the size and number of their wheels, ferry operators by their lengths, and customs officials by their contents. Once the purpose of

the classification has been clearly specified, the taxonomist can set about choosing and weighting the characters he considers appropriate, though, as we mentioned above, if the specification is not sufficiently comprehensive, different taxonomists may have different views about the characters.

The third type of taxonomy involves what Sneath and Sokal (1962) call "phylogenetic speculation". We may refer to it as phylogenetic taxonomy, and note that it does not primarily involve classification, since its avowed purpose is the estimation of the evolutionary tree which supposedly unites the organisms being studied. Characters must be chosen and weighted according to their evolutionary significance, and judgement of this is likely to be subjective. However, the more that is known about the genetics of the organisms, the easier it will be to agree about the relevance of characters. In Pavia over the past two years we have studied in considerable detail the problem of the statistical estimation of the form and dimensions of evolutionary trees, and we describe our current approach below. The logical basis of such studies seems to be rather simple, probably because their purpose can be so simply and clearly defined, but it is important to realise that the procedure is not primarily classificatory, although, once an estimate of the form of an evolutionary tree has been produced, organisms can of course be grouped according to relative evolutionary divergence.

Our three-taxon classification of taxonomy, and our belief that it holds the key to a satisfactory logical structure of taxonomic studies, will, no doubt, be criticized. It may well be asked, for example, whether the classification of sheep as sheep and goats as goats * is pure, applied, or phylogenetic in type, and, if applied, what is the specific purpose of the classification? Is it not really a general purpose classification? Our answer to this is that the classification *was* indeed applied, but that the original division into sheep and goats was for a purpose which is lost in antiquity. However, it was found that the division was useful in other ways, or, to look from a different point of view, classification of the animals for other purposes frequently led to the same division. The division was evidently rather stable, so much so that it was appropriate to name the animals in the two classes as sheep and goats. Such a classification is acceptable to us for two main reasons: first, because sheep and goats are really rather different, and intermediate forms do not exist, so that there is never any doubt as to whether an animal is a sheep or a goat. The subject matter of the classification is 'robust' with respect to the different purposes to which the classification is to be put. Only robust classes of objects acquire names. But secondly, and more important, the classification is acceptable because we understand its limitations. We know that classification into sheep and goats is not much use if we are interested in the proportion of the combined flock that is male, or the proportion that is under a certain age. In fact

* The terms "sheep" and "goat" refer here to the West European domesticated races of *Ovis aries* L. and *Capra hircus* L., respectively.

our classification is not really general purpose at all; though it arose out of an applied taxonomy, because of the stability of the classes, it has become a pure taxonomy, descriptive in intent. When we say "sheep" we mean, according to the Oxford Dictionary, a "timid gregarious beast kept in flocks for the mutton or lamb and wool and leather it yields". A goat, on the other hand, is "a lively wanton strong-smelling usually horned and bearded ruminant quadruped".

Our present problem, however, is not that of considering how classifications arose historically, because we are trying to set up a classification which is logically acceptable, and does not rely on age-old decisions. It is as though we were confronted with a field full of sheep and goats for the first time, and were asked to classify the animals. Our reaction must be to ask to what purpose the classification will be put, or, alternatively, to ask what are the characters that we should consider. Thus we can embark on an applied or a pure taxonomy, whichever is appropriate. But if we are asked to classify the animals "for general purposes" we must insist on knowing *what* purposes, though, of course, there is nothing wrong in admitting that the animals seem at first sight to be divisible into two homogeneous classes, and that this classification will no doubt serve the majority of purposes included under the title "general". *But* this fact derives from the particular data we are using, and cannot be used as a justification for a taxonomic procedure whose logical structure must be independent of the organisms being considered. The fact that different characters and different taxonomic methods lead to the same segregation of the animals into sheep and goats tells us something about sheep and goats, but nothing about the logical basis of taxonomy.

3. Reconstruction of Evolutionary Trees

Having detached phylogenetic taxonomy from purely classificatory studies, we may now proceed to examine it more closely. The characters which are fundamental to differentiation are the base-sequences of DNA. These are at present unavailable directly, although methods have recently been suggested to measure the extent to which two DNA's match each other for base sequence. In the absence of such data, use must be made of genotypic, or even phenotypic, differences. At the genotypic level the relative importance of various gene substitutions in evolution is a matter which only detailed study can reveal; it will depend upon the past and present selective forces, mutation rates and population structures and sizes. At the phenotypic level such information is even less readily available.

As has been found with classification, it is convenient to represent the variation amongst the organisms geometrically, the chosen characters being represented by axes in a multidimensional space, and the organisms, specified by these characters, being represented as points in this hyperspace. If a time dimension, everywhere normal to the character space, is added, the course of evolution — were it but known — could be seen as a 'tree', sometimes branching as groups of organisms diverged, sometimes growing together as hybridization took place, and

with many branches ending in extinction before the remainder intercepted the 'now' character space to indicate the current disposition of the organisms. This approach presupposes that the character-space is continuous, which will normally be an acceptable approximation. In cases in which discreteness is apparent the space will take the form of a lattice of points.

The proper basis for a study of evolutionary divergence will be provided by means of that transformation of the space-time which, as far as is known, makes a unit vector, in whatsoever direction (normal to time) and in whatsoever part of the space-time, correspond to a fixed amount of evolutionary divergence — divergence being properly defined for the problem under consideration. Such a transformed space-time would thus be everywhere isotropic with respect to evolutionary progress. Transformation is, in this context, equivalent to character-weighting.

Put thus formally, it will be obvious that the knowledge necessary for the reconstruction of evolution is rarely available, but, as in other problems of statistical estimation, deductions can always be made provided the assumptions on which they rest are remembered.

The present problem is to estimate the form of the evolutionary tree given only the information contained in the 'now' character-space. Fossils, which would appear at other time levels in space-time, will not be treated at present, although we would like to stress particularly that it is clear from this approach that there is no substantial logical difference about estimating the course of evolution with or without fossil evidence.

In our isotropic space we may think of evolution as a branching random walk, with a constant probability of branching and a constant rate of walking: that is, after the elapse of a certain time interval, the probability distribution in space of the position of a population will be normal, with mean at its original position and variance proportional to the time elapsed. The constant of proportionality will depend on many parameters, such as the population size and the type and intensity of selection, and will usually be unknown. Since an evolutionary tree uniting n points (without loops) is bound to contain $n-1$ branching points, the probability of branching is not a parameter which has to be estimated, although it is worth noting that the theory of birth and death processes may have some interesting things to say about the expected form of an evolutionary tree.

Now the probability density at a distance d spatially and t temporally from a point in the space-time of p spatial dimensions is given by

$$\sigma^{-p}\,(2\pi t)^{-\frac{1}{2}p}\exp\{-d^2/(2t\sigma^2)\}$$

where σ^2 is the constant of proportionality mentioned above. The log-likelihood is therefore given by

$$-\{d^2/(2t\sigma^2) + \tfrac{1}{2}p\log(2t\sigma^2) + \tfrac{1}{2}p\log\pi\}$$

Writing T for $2t\sigma^2$, omitting the constant, and changing the sign, the expression becomes

$$d^2/T + \tfrac{1}{2}p \log T$$

Each arm of a postulated evolutionary tree will have an expression of this type associated with it, d being its spatial length and T proportional to its temporal length. Thus the likelihood of the tree will be maximized if the sum of these expressions, over all the arms, is minimized (since we have changed the sign, and the branching points, being constant in number, are irrelevant).

We have found that the mathematics of the estimation procedure can be made quite compact using a matrix notation, and we do not anticipate any major difficulties in the corresponding computer programmes, although these are not yet working. However, as will be mentioned below in connexion with our example, other methods have been used which may be expected to give similar results.

Thus maximum-likelihood estimation is not difficult once the topological form of the tree has been specified, but unfortunately there will generally be too many forms for the maximum likelihood of each to be evaluated: with n points to unite there are $(2n-5)!/(n-3)!\ 2^{n-3}$ trees, or more than two million with ten points. We therefore have to resort to some prior method of clustering. Initially we developed our own method, based on the analysis of variance (Edwards and Cavalli-Sforza, 1963b), but we have finally adopted a method which is particularly simple and rapid. Prim (1957) has shown that the network of minimum length uniting n points, but with the segments constrained to meet only at these given points, can be found by listing all the distances between points in increasing order, and successively allocating segments to these distances, omitting any segment which completes a loop. The resulting net gives some indication of the types of tree it will be worthwhile to use in the maximum-likelihood estimation procedure. The relationship is somewhat intuitive at the moment, but once one is "on the track" of a good topology, experience shows what sort of changes will be likely to increase the overall likelihood. In particular, if the length of an arm is estimated as zero, a change in the topology of that part of the tree is indicated.

We are now left with the problem of finding the correct transformation of the character space. The transformation is, of course, a reflection of the genetic assumptions which are being made, and these will be peculiar to each case. For the moment it will be sufficient to indicate how we have proceeded in a particular example.

4. The Recent Evolution of Man

(a) *Materials*

As our example, we decided to study the phylogeny of fifteen samples of human populations, using as characters the frequencies of various blood-group alleles. This choice was made because of the availability of the data, the intrinsic interest of the problem, and the amount of

knowledge about the genetic situation. Data on the blood-group systems A_1A_2BO, Rh (four sera), MNSs, Fy and Di were obtained for fifteen populations, three from each continent, with the invaluable help of Dr. Mourant and Mrs. Sobczak, of the Medical Research Council Blood Group Reference Library, London. In all, twenty 'genes' were distinguished. Nearly every sample consisted of more than 70 individuals, and was tested for each of the five groups; although there is a certain amount of heterogeneity with respect to the sera used, and the methods of calculating the gene frequencies, these data were considered adequate for a trial of the methods.

The five blood groups were assumed to be independent, because of the absence of evidence for linkage between them. Each allele within a system, and each of the systems, was given equal weight. In view of the apparent absence or weakness of selective forces in these blood groups, and, in cases where differential mortality has been demonstrated, of the impossibility of setting up selective models of sufficient validity, it has been assumed for our preliminary work that selective forces are absent, and that the observed divergences are, or can be treated as being, due to drift. Boyd's comment (Boyd, 1963), that "unless the blood groups are adaptive, they are not going to be very useful in racial classification", is not acceptable: any variable, even a random one, which shows a persistently high correlation between successive generations, may be useful. If differential selective forces are operating, the simultaneous use of several independent loci should render their effects on the analysis minimal. All the populations have been treated as though they were, and always have been, of the same size and structure.

(b) *Methods*

It is now necessary to set up a character-space on the basis of these assumptions, following the methods outlined above. Since the normal scale of gene frequencies is anisotropic in the sense that the effects of drift are differently measured in different parts of the scale, a transformation is necessary. The 'square-root' transformation given by

$$\sin^2\theta = p, \qquad \cos^2\theta = q,$$

is appropriate for two alleles at a locus (Fisher, 1954), since it stabilizes the binomial variance throughout the range, and Fisher (personal communication) has pointed out that it may be extended to the case of many alleles. With n alleles, populations will be represented as points on $(1/2^n)$th of the surface of the unit hypersphere in n dimensions, the angular distance between populations with gene frequencies $(p_1, q_1, r_1, \ldots)$ and $(p_2, q_2, r_2, \ldots)$ being given by

$$\mathrm{Cos}\alpha = \sqrt{p_1p_2} + \sqrt{q_1q_2} + \sqrt{r_1r_2} + \ldots$$

Since the character-space is curved, it will be necessary to transform it into a Euclidean space before the available maximum-likelihood treatment can be applied. As in map projections, some desirable qualities of the attribute-space will be lost in the transformation. The most

appropriate transformation is worth a separate study, but the simplest approach is to use, as the measure of distance between two points, the chord — the Euclidean straight line — rather than the arc of the great circle on the surface of the hypersphere. This is equal to $\sqrt{(2-2\cos\alpha)}$. The maximum possible error thus incurred is given by the ratio of the length of the arc of a quadrant of a circle to that of the corresponding chord, or $\sqrt{2} : \frac{1}{2}\pi$, amounting to a maximum underestimate of distance of 10%.

Given the Euclidean distance between two populations for each locus, the total distance over all loci is found by taking the square root of the sum of the squared distances for individual loci, by Pythagoras' theorem, since the character-spaces for the loci are mutually orthogonal. The resulting array of pairwise distances between populations is ready for finding the Prim network. In order to use the maximum-likelihood computer program it is necessary to set up a system of Cartesian co-ordinates for the points representing the populations, and this may be done on a computer from the pairwise distances. The space thus created is the final version of the character-space, but with an arbitrary frame of reference. The unit distance, however, is not arbitrary, and corresponds to an amount of evolution equal to that incurred in a single gene substitution at a single locus.

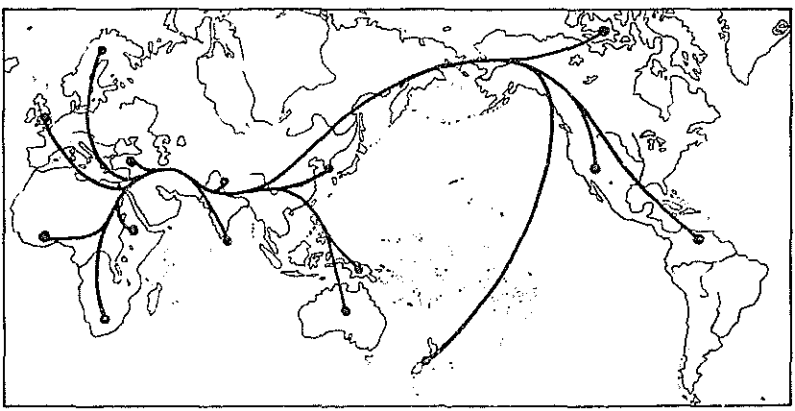

FIG. 1. Topology of the minimum-evolution tree uniting fifteen human populations; constructed on the basis of the frequency of blood-group alleles.

Unfortunately, as has been mentioned, our maximum-likelihood programmes are not yet working, but, using this character-space, our earlier "method of minimum evolution" (Edwards and Cavalli-Sorza, 1963a) gave a 'best' tree of the topological form shown in the figure (Fig. 1). It is probable that this method gives a tree which is approximately the same as the projection of the maximum-likelihood tree onto the 'now' character space. Whilst we would be the first

to admit that we have been fortunate to get such a reasonable result from preliminary data, we feel that it is at least encouraging, and justifies our intention to develop our methods further, and apply them to more extensive bodies of information.

5. Acknowledgements

This work has been supported by a grant from the U. S. Atomic Energy Commission and by Euratom-CNR-CNEN contract no. 012-61-12 BIAI.

REFERENCES

Boyd, W. C., 1963. Genetics and the human race. *Science*, **140**: 1057-1064.

Edwards, A. W. F., and Cavalli-Sforza, L. L., 1963a. The reconstruction of evolution. Unpublished paper read at 142nd meeting of the *Genetical Society of Great Britain*, London, July 1963. (Abstract in *Ann. hum. Genet.*, **27**: 104-105, and in *Heredity, Lond.*, **18**: 553.)

—— —— 1963b. A method for cluster analysis. Unpublished paper read at *5th Int. Biometric Conf.*, Cambridge, 1963.

Fisher, R. A., 1954. *Statistical Methods for Research Workers*, ed. 12. Oliver & Boyd, Edinburgh. 356 pp.

Maccacaro, G. A., 1958. La misura della informazione contenuta nei criteri di classificazione. *Annali Microbiol.*, **8**: 231-239.

Prim, R. C., 1957. Shortest connection networks and some generalizations. *Bell Syst. tech. J.*, **36**: 1389-1401.

Resigno, A., and Maccacaro, G. A., 1961. The information content of biological classifications. In: C. Cherry (ed.), *Information Theory*, 437-446. Butterworth and Company, London.

Sneath, P. H. A., and Sokal, R. R., 1962. Numerical taxonomy. *Nature, Lond.*, **193**: 855-860.