Phylogenies and the Forces of Evolution

FRANK B. LIVINGSTONE
Department of Anthropology, University of Michigan, Ann Arbor, Michigan 48109

ABSTRACT

The construction of phylogenetic trees from gene frequency data assumes that a history of binary fissioning of populations has been the major cause of genetic variation. However, in many areas of the world human populations have been relatively stable with local gene flow. This population history is closer to an isolation by distance model. It was modelled by a simulation of gene frequency changes in a linear sequence of 50 stable populations with gene flow among neighboring populations. Phylogenetic trees were constructed from the gene frequencies after the simulation was run for 500 generations. Using only a few loci there is little correlation between genetic and geographic distance, but with 40 or more loci, there was a perfect correlation with geographic distance. A different population model can thus result in a phylogenetic tree comparable to those assumed to be produced by binary fission.

In the last few years there has been a continuing and literally explosive increase in knowledge of genetic variation among human populations. As would be expected, the understanding or explanation of this variation has lagged behind the pace of discovery. One reason is that the analysis of such variation in terms of the forces of evolution has not been a major concern; instead most efforts are attempts to use the genetic variation to reconstruct human demographic history, primarily by the construction of phylogenetic trees. The tree for a single locus or closely linked loci frequently results in a phylogeny that is obviously wrong, but the current view is that if enough loci are used, such errors will even out and so produce a “true” tree. Cavalli-Sforza et al. (1988) is the most comprehensive recent attempt that uses most known genetic variation to construct a phylogenetic tree for the human species. Commenting on this work, Gould (1989: pp. 22–23) has asserted, “The best way to work past these difficulties lies in a ‘brute force’ approach: the greater the quantity of measured differences, the greater the likelihood of a primary correlation between time and overall distance.” Gould’s brute force approach does have a theoretical basis (Nei et al., 1983), but other models of population history in addition to phylogenetic trees can result in a similar pattern of genetic variation.

The basic assumption of this approach is that genetic change is primarily due to neutral evolution. All the differences in gene frequencies that are used to construct the phylogeny are assumed to be due to genetic drift and possibly gene flow. Nei et al. (1983) have shown that one or a few loci will give an inaccurate estimate, and the error decreases gradually, so that with about 50 loci the estimate approaches the actual populational phylogeny with minimum error. More loci do not increase the accuracy very much. The same problem exists for the construction of species trees (Pamilo and Nei, 1988). The human data used by Cavalli-Sforza et al. (1988) with 42 loci and 120 alleles are approaching this necessary amount, and it is interesting to note that the original phylogenetic tree constructed by Cavalli-Sforza et al. (1964) on much less data was very different, with Europeans more closely linked to Africans than to Asians. It is also noteworthy that neither study used the great amount of data known for the hemoglobin and G6PD loci because of the known selection operating at these loci. However, the A1A2BO, FY, HLAA, HLAB, and El loci are also known to have some selection operating on them and are used in their most recent tree construction. Other recent studies using hemoglobin restriction site polymorphisms to construct phylogenetic trees (Wainscoat et al., 1986; Long et al., 1990) make the same assumption of neutrality.

In addition to neutral evolution, phyloge-
Phylogenetic tree construction also assumes that the differences are caused by population fission. There have been many reports that find strong correlations between linguistic, geographic, and genetic differences; the detailed analysis of Barrantes et al. (1990) of several central American tribes is the latest on a small region, while Sokal (1988) has reported on correlations throughout Europe. These correlations presumably result from the same causes of differentiation of language and genes. Cavalli-Sforza et al. (1990: p. 18) have stated, "... the two evolutions follow in principal the same history, namely sequence of fissions. Two populations that have separated begin a process of differentiation of both language and genes." Nevertheless, other processes can result in the differentiation of populations. Phylogenetic trees are only one way of describing differences among a set of populations or entities, and geographic variation can result from many other forces and very different demographic histories. The purpose of this paper is to show that other processes or a very different demographic history of these populations will result in phylogenetic trees that have an almost perfect correlation with geography and thus seem to be equally plausible explanations of genetic variation among humans.

Clinal variation can be due to other forces, especially natural selection, and will produce a "good" phylogeny that seems to accord with other data such as linguistic or geographical variation providing the cline is monotonically increasing or decreasing. Variation in hemoglobin S and blood group A genes in Liberia is shown in Figure 1. The phylogeny based on the hemoglobin S cline is shown in Figure 2. I have not found any cognate data to measure linguistic differences, but there appears to be some clustering of closely related languages having similar gene frequencies. The correlation with geographical distance is .55. The most reasonable explanation for this cline is the wave of advance of an advantageous gene and demonstrates that clines that are due to other forces do produce good phylogenies. If the blood group A data are added, the resulting tree is a poorer fit with language and geography (Fig. 3), and adding five more blood group alleles does not improve the tree very much. A combination of many loci would increase the probability of including other loci with variation approximating a monotonic cline, and these would seem to account for the convergence to a perfect tree.

The analysis of most clines whether they are stable or advancing assumes that the populations are reasonably stable and population displacement or expansion are not occurring, although a cline could also be due to population expansion. If we assume that the human species has occupied much of the area of the Old World for a long time and most gene flow has been among neighboring populations, then phylogenetic trees can be constructed even in the absence of population fission. This is essentially the isolation by distance model (Cavalli-Sforza and Bodmer, 1971), where most of the differentiation of the populations is due to gene drift and restricted by local gene flow. To model this, a linear series of 50 populations were all begun with the same gene frequencies of .5 for two alleles. After 500 or 1,000 generations the variation or cline was recorded. Using the same initial frequencies in all populations is comparable to assuming that the region was occupied by a rapid population expansion in a relatively short time and then has remained relatively stable. Thus, the genetic variation has been produced by isolation by distance. Migration rates of .05 with adjacent populations, .01 with the two populations adjacent to the adjacent ones, and .01 with a population randomly drawn from five on either side were assumed to be a reasonable approximation to early humans. The rate of .05 had been used by others (Weiss and Maruyama, 1976; Rouhani, 1989), but I have found that occasional long-distance mi-

![Fig. 1. The clines for the hemoglobin S and blood group A genes in Liberia.](image-url)
移民可以对基因扩散产生重要影响（Livingstone，1989）。每个群体被设置为每代200个个体或400个基因，尽管每次选择程序在某些运行中减少了每代的群体数量，其他群体大小被使用，图4和5是500代无选择或突变后形成的两个例子的种群，即群体多样性在500代后没有显著变化。显然，任何一次运行，模拟一个等位基因，不会给出与地理距离相关联的树，即使使用10个等位基因，许多群体也被认为“不在合适的位置”如图6所示。然而，50个等位基因的完美拟合与地理距离相结合（图7）。图中显示的是每三个群体或每五个群体的集群分析，但结果相同，并且在所有情况下，50个等位基因均能产生完美树。
Fig. 4–5. Clines generated after 500 generations with .05 gene flow between adjacent populations, .01 with those two removed, and .01 gene flow randomly assigned to a population within 5 on either side.
Fig. 6–8. See overleaf for legend.
Fig. 6-10. The phylogenetic trees constructed from the simulation of a linear sequence of 50 populations with various values of the parameters as indicated.

Analysis used was the UPGMA options of the MIDAS program developed at the University of Michigan.

Using a smaller population size does not affect the result. If selection is programmed to act identically in all populations, then as little as .05 selection against both homozygotes seems to increase at times the number of loci needed to produce a perfect tree.

In order to determine the effects of mutation and the generation of new genetic variation, an infinite alleles model was programmed among an identical set of 50 populations with the same amount of migration. Figure 8 shows that with ten loci there will be many populations out of place, but with 60 loci there will be a perfect fit (Fig. 9). On other runs a perfect fit was obtained with as few as 40 loci. In these runs with no selection, there was an average of 3 to 4 alleles at each locus after beginning with two alleles at .5 each and a mutation rate of .00001. This model was also run for 1,000 generations and there did not seem to be any increase in the number of alleles present, and again a perfect tree was produced. Selection of .05 against homozygotes was added, but it did not change the results as shown in...
Figure 10. With this model, selection of even .1 did not prevent a perfect tree but did increase the numbers of alleles present, which seems plausible since selection against homozygotes would tend to favor rare alleles.

These results are not unexpected. Similar problems and models were discussed by Felsenstein (1982), Harpending and Ward (1982), and many others. The results show that genetic variation that has a significant correlation with geography can be explained by a number of models other than binary fission. Life is not entirely binary fission. Harding and Sokal (1988) have also concluded that short-range interdemic gene flow is a major cause of the genetic and linguistic variation in Europe. Although Barrantes et al. (1990) interpret their data as due to phyletic fissioning, they also point out that the populations of Chibcha Amerindians have probably inhabited the same region for almost 10,000 years and that adjacent populations tend to resemble each other. I do not think they have excluded the model used in this paper of local migration as the cause of genetic variation and of the correlation of genetic and geographic distance.

It has seemed to me for years that the understanding of genetic variation is best approached one locus at a time, and that instead of a "brute force" and ignorance approach, an analysis of how the clines for a specific locus can be produced by the forces of evolution could lead to a better understanding of genetic variation. Among the rather small isolates in the "underdeveloped" world, the genetic variation will be due primarily to a balance of gene flow and gene drift, and a consideration of several loci would lead to a better estimate of these forces. However, the distributions of most polymorphic loci among the world's populations must surely be due in part to other forces.

LITERATURE CITED


