Commentary

Bhattacharyya's distance measure as a precursor of genetic distance measures

In Population Genetics, two populations are distinguished from each other on the basis of the differences in the distributions of the alleles at the locus or loci under consideration. These differences are measured by a "genetic distance" between the two populations (not to be confused with genetic distance between two loci, which is based on recombination fractions) and they play a major role in inferences at the population level. Several measures of genetic distance have been proposed by different authors (Sanghvi 1953; Cavalli-Sforza and Edwards 1967; Jukes and Cantor 1969; Nei 1972; Kimura 1980; Reynolds *et al* 1983; reviews in Felsenstein 1991; Nei and Kumar 2000). Most of these measures are actually dissimilarity measures and not mathematically true distance measures (B-Rao and Majumdar 1999). Independently, and much before the geneticists, statisticians too were concerned with the idea of distinguishing between two (statistical) populations. In order to discriminate between two populations on the basis of one or more characters, divergence measures like "Mahalanobis' D^2 statistic" or "Mahalanobis' generalized distance" (1936) and "Bhattacharyya's distance" (1943, 1946), Kullback-Leibler's divergence measure (1951) etc. have been proposed by statisticians. Mukherjee and Chattopadhyaya (1986) have mentioned measures based on distances, association between two attributes and discrimination function. There are similarities between the distance measures defined by applied scientists and by theoreticians. Felsenstein (1985) shows that three of the allele frequency-based genetic distance measures were anticipated by Bhattacharyya (1946). Nei and Takezaki (1994) have also studied the effectiveness of several genetic distance measures in the context of phylogenetic analysis, including Bhattacharyya's distance measure.

1. Bhattacharyya's distance

Bhattacharyya (1946) defined a measure of distance between two populations, based on the number of occurrences (counts) of each of k traits. An individual in the population possesses exactly one of the k traits, or in other words, falls into exactly one of k classes. Suppose the i'th trait occurs n_i times in a population, $i = 1, 2, \ldots, k$. Then, one can define the population with respect to these k traits, by the profile given by, $\underline{n} = (n_1, n_2, \ldots n_k)$; $\Sigma_i^k n_i = N$, where N is the population size. Then $\underline{n}$ follows multinomial distribution, which is a generalization of the binomial distribution to k classes ($k > 2$). Such a statistical population is called a multinomial population. These counts can be converted to frequency or probability distributions.

Consider two multinomial populations with probability distributions $(\pi_1, \pi_2, \ldots, \pi_k)$ and $(\pi_1', \pi_2', \ldots, \pi_k')$ where $\Sigma\pi_i = \Sigma\pi_i' = 1$. Geometrically, these distributions can be plotted as points in k-dimensional space by taking $(\sqrt{p_1}, \sqrt{p_2}, \ldots, \sqrt{p_k})$ and $(\sqrt{p_1'}, \sqrt{p_2'}, \ldots, \sqrt{p_k'})$ to be the direction cosines of two straight lines through the origin. If Δ is the angle between these two lines, then

$$\mathrm{Cos}\Delta = \sum_{i=1}^{k} \sqrt{p_i p_i'} \,. \tag{1}$$

The square of the angle between these two lines can then be taken as a measure of divergence between the two populations.

$$\Delta^2 = \left(\mathrm{Cos}^{-1} \left(\sum_{i=1}^{k} \sqrt{p_i p_i'} \right) \right)^2 . \tag{2}$$

This measure was proposed by Bhattacharyya (1946) and is known in the statistics literature (as also in some application areas like physics and computer science) as Bhattacharyya's distance.

Bhattacharyya also defined a statistic Δ' similar to Δ, based on sample proportions, $\{p_i\}$ and $\{p'_i\}$ for the two samples. However, he showed that Δ'^2 is not an unbiased estimator of Δ^2. Geometrically, he interpreted Δ' as a rectilinear (or chord) length, joining the two sample points located at the intersection of the two sample lines with the unit hypersphere in k-dimensional space.

$$\mathrm{Cos}\Delta' = \sum_{i=1}^{k} \sqrt{p_i p'_i}\ . \tag{3}$$

In population genetics, a population is defined by the frequencies of alleles at a given locus or a set of loci. For a single locus, the analogy to a multinomial population is obvious and the above distance measure is directly applicable. However, when the population is defined by its state at a set of loci, this distance measure can be applied after generalization by summing over all loci.

2. Population genetic distances

Numerical taxonomists (Sneath and Sokal 1973) have sought to draw inferences about the grouping of taxa (families or genera or species or populations) based on the similarities (equivalently, dissimilarities) between extant taxa. Many different measures of similarity or dissimilarity have been defined by them. Taking off from these pheneticists, early phylogeneticists also sought to draw evolutionary inferences between different taxa based on the similarities between the extant taxa. Several measures of population genetic distance based on molecular data have been proposed (Jukes and Cantor 1969; Kimura 1980; Felsenstein 1991). One set of genetic distances is based on allele (or isozyme) frequency distributions at one or more genetic loci (Cavalli-Sforza and Edwards 1967; Nei 1972; Reynolds *et al* 1983). Of these, the ones that are self-evidently closest to Bhattacharyya's work are the Cavalli-Sforza distance measures.

Like Bhattacharyya, Cavalli-Sforza and Edwards also thought of placing the square roots of (allele) frequency distributions as points in k-dimensional space. They attribute this generalization from 2 alleles (Cavalli-Sforza and Conterio 1960) to k alleles, to a personal communication with Fisher (Cavalli-Sforza and Edwards 1967).

They also considered the angular distance q between two populations, which is the same as the Δ of Bhattacharyya. However, in order that unit distance represent one gene substitution, they recommended using as a measure of distance (q measured in radians),

$$\Delta_{arc} = 2q/\pi. \tag{4}$$

This measure is known as the Cavalli-Sforza arc distance, for, the length of the arc of the unit hypersphere between the two points representing the two populations is given by q. They noted that the points would lie only on the surface of the $(1/2^k)$th part of the unit hypersphere which is positive in all co-ordinates (for example, the surface of the positive octant of a sphere in 3-dimensions). Their chord distance is defined as (for single locus),

$$\Delta_{ch} = (2\sqrt{2}/p)\sqrt{1-\mathrm{Cos}q}\ , \tag{5}$$

where

$$\mathrm{Cos}q = \sum_{i=1}^{k} \sqrt{p_i p'_i}\ ,$$

and π_i, π'_i are frequencies of the ith allele in populations 1 and 2 respectively, and $2/\pi$ is a scaling constant. The similarities between Bhattacharyya's distance [eqn (2)], the arc distance [eqn (4)] and chord distance [eqn (5)] are evident.

The underlying concept behind the Cavalli-Sforza's distances is the same as that of Bhattacharyya's distance – geometrical placement of the frequency distributions as points in multidimensional Euclidean space, direction cosines, angular distance and chord length.

Equally independently of Bhattacharyya, Sanghvi (1953) defined a measure of distance betwen two populations based on attribute (qualitative classes) data, which is analogous to the c^2 distance. In Sec. 9, page No. 405 of Bhattacharyya's (1946) paper, he derives a formula (which he attributes to Mahalanobis's suggestion), which is very similar to Balakrishnan and Sanghvi's distance (1968). As pointed out by Felsenstein (1985), although the derivations are different, the final formulas are very similar.

3. Has Bhattacharyya been neglected by geneticists?

We attempted to scan the phylogenetic analysis literature for references to Bhattacharyya's paper in the context of genetic distances. Many authors still seem to be unaware of Bhattacharyya's paper. Even those who are aware of it have cited it but rarely.

Sanghvi (1953) was apparently unaware of Bhattacharyya's (1946) work. Cavalli-Sforza and Edwards (1967) were unaware of both Bhattacharyya's (1946) and Sanghvi's (1953) work. Balakrishnan and Sanghvi (1968) mention Cavalli-Sforza and Edwards (1967), but not Bhattacharyya (1946). It appears that statistical geneticists were ignorant of Bhattacharyya (1946) till about 1985. The Memorial meeting in honour of Prof. Bhattacharyya (Mukherjee *et al* 1994) has gone some way in drawing attention to his work, but fails to focus on the significance of his distance measure in applied areas, particularly in genetics.

Felsenstein (1985) mentions the connection between Bhattacharyya's distance with each of Cavalli-Sforza's arc distance, Cavalli-Sforza's chord distance, and Balakrishnan and Sanghvi's distance (1968). However, although the documentation of the highly popular PHYLIP (Felsenstein 1991) software package contains a very comprehensive bibliography, it does not include Bhattacharyya's 1946 paper. Felsenstein referred to the chord distance as the "Cavalli-Sforza chord distance" in his PHYLIP package, and it continues to be known by that name. Cavalli-Sforza *et al* (1994, p. 29) admit that the angular transformation is equivalent, for a single locus, to a formula by Bhattacharyya (1946).

Weir (1996) has discussed geometric measures of genetic distance at some length. Although he does not state this explicitly, he has shown that Nei's standard genetic distance can also be derived using geometric reasoning (p. 192, Weir 1996); it can be interpreted as the cosine of the angle between the lines joining the points representing the two populations with the origin [equivalent to eqn (1) above]. He also discusses the distance between a population and a sample drawn from it, and reproduces the derivation of eqns 3·1, 3·2 and 3·3 in Bhattacharyya (1946). However, there is no mention of Bhattacharyya (1946), who preceded Weir by half a century. Weir (p. 194, 1996) also derives the Balakrishnan and Sanghvi distance (1968) without citing the prior work.

Nei and Takezaki (1994) have mentioned Bhattacharyya's distance Δ^2 (which they call q^2), which, in fact, they found to be one of the best for correctly estimating the topology of a phylogenetic tree. Nei and Kumar (2000) cite Bhattacharyya's angular transformation and its modifications in order to give the expression for the distance q^2 between two populations. However we found no mention of Bhattacharyya's paper in Hillis *et al* (1996), which is a major reference work for anyone doing research in the field of molecular systematics.

Bhattacharyya has not found his place in genetics, but his distance measure has been applied and mentioned elsewhere. A simple search of the Internet extracts at least 203 references to Bhattacharyya's distance. Bhattacharyya's distance has found wide application in computer science (for example, in B-fitting in machine learning, face recognition), physics and ecology. XLSTAT News (22 January 2003) refers to inclusion of similarity/dissimilarity and hierarchical clustering on the basis of Mahalanobis's distance, and Bhattacharyya's distance (www.xlstat.com/news.htm), indicating its growing importance.

4. Conclusion

Bhattacharyya (1946) cites the influence of Mahalanobis (1930) while defining his measure of diversity between two populations. His concept of viewing frequency distributions as points in geometric space has found wide applications in diverse fields, including genetics. Evolutionary geneticists routinely use various distance measures, like Nei's standard genetic distance, Cavalli-Sforza's arc or chord distance and Balakrishnan and Sanghvi's distance, all of which were explicitly or implicitly contained within Bhattacharyya's work. However, Bhattacharyya's work, which preceded the others by two to five decades, is rarely cited in any of the most prominent and visible works in phylogenetic analysis. It is only appropriate that the record be set right by researchers in the field.

References

Balakrishnan V and Sanghvi L D 1968 Distance between populations on the basis of attribute data; *Biometrics* **24** 859–865

Bhattacharyya A 1943 On a measure of divergence between two statistical populations defined by their probability distributions; *Bull. Cal. Math. Soc*. **35** 99–110
Bhattacharyya A 1946 On a measure of divergence between two multinomial populations; *Sankhya* **7** 401–406
B-Rao C and Majumdar K C 1999 Reconstruction of phylogenetic relationships; *J. Biosci*. **24** 121–137
Cavalli-Sforza L L and Conterio F 1960 *Atti. Assoc. Genet. Ital*. **5** 333–344 (in Italian)
Cavalli-Sforza L L and Edwards A W F 1967 Phylogenetic Analysis: Models and estimation procedures; *Evolution* **32** 550–570
Cavalli-Sforza L L, Menozzi P and Piazza A 1994 *The history and geography of human genes* (Princeton: Princeton University Press)
Felsenstein J 1985 Phylogenies from gene frequencies: A statistical problem; *Syst. Zool*. **34** 300–311
Felsenstein J 1991 PHYLIP (Phylogeny inference package) v.3·4. University of Washington, Seattle, USA
Fitch W and Margoliash M M 1967 Construction of phylogenetic trees; *Science* **155** 279–284
Hillis D M, Moritz C and Mable B K (eds) 1996 *Molecular systematics* (Sunderland: Sinauer Associates)
Jukes T H and Cantor C R 1969 Evolution of protein molecules; in *Mammalian protein metabolism* (ed.) H N Munro (NewYork: Academic Press) pp 21–132
Kimura M 1980 A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences; *J. Mol. Evol*. **16** 111–120
Kullback S 1997 *Information theory and statistics* (New York: Dover Publications)
Kullback S and Leibler R A 1951 On Information and Sufficiency; *Annals of Math. Stat*. **22** 76–86
Mahalanobis P C 1930 On the tests and measures of group divergences; *J. Proc. Asiatic Soc. Bengal* (*New Series*) **26** 541–588
Mahalanobis P C 1936 On the generalised distance in ststistics; *Proc. Natl. Inst. Sci. India* **2** 49–55
Mukherjee S P and Chattopadhyaya A K 1986 Measures of mobility and some associated inference problems; *Demography India* **15** 269–280
Mukherjee S P, Chaudhuri A and Basu S K (eds) 1994 *Essays on probability and statistics* (*Festschrift in Honour of Professor Anil Kumar Bhattacharyya*) (Calcutta: Department of Statistics, Presidency College)
Nei M 1972 Genetic distance between populations; *Am. Nat*. **106** 283–292
Nei M and Kumar S 2000 *Molecular evolution and phylogenetics* (Oxford: University Press) pp 266
Nei M and Takezaki N 1994 Estimation of genetic distances and phylogenetic trees from DNA analysis; *Proc. 5th World Congr. Genet. Appl. Livestock Prod.* **21** 405–412
Reynolds J B, Weir B S and Cockerham C C 1983 Estimation of the co-ancestory co-efficient; basis for a short term genetic distance; *Genetics* **105** 767–779
Sanghvi L D 1953 Comparison of genetical and morphological methods for a study of biological differences; *Am. J. Phys. Anthrop*. **11** 385–404
Sneath P H A and Sokal R R 1973 *Numerical taxonomy: The principles and practice of numerical classification* (San Francisco: W Freeman)
Weir B S 1996 *Genetic data analysis 2: Methods for discrete population genetic data* (2nd edition) (Sunderland: Sinauer Assoc.)

APARNA CHATTOPADHYAY,
ASIS KUMAR CHATTOPADHYAY*,
CHANDRIKA B-RAO†,
Genome Informatics,
Institute of Genomics and
Integrative Biology,
Mall Road, Delhi 110 007, India
**Department of Statistics,*
University of Calcutta,
35, BC Road,
Kolkata 700 019, India
†(Email, cbrao@igib.res.in)