Applications of Genetic Distances on Blood-group Gene Frequencies and their Statistical Genetic Similarities

Abstract

This paper studied the applications of genetic distances on Blood-group gene frequencies and their Statistical genetic similarities characterizing four populations, for data from Eakimo, Bantu, English and Korea. The study compared the following distances: Euclidean, squared Euclidean, Minkowski, Chebychev and City Block on the above mentioned data for the outlined countries. Correlation analysis was applied to evaluate the relationships between these countries on their blood group gene frequencies. Similarity check was also conducted to know the countries that have similar blood-group gene frequency. It was observed that Euclidean distance and Minkowski distance had equal distances. This means that the two distances are more similar in this particular data set than the other studied distances. The study revealed that Chebychev distance had the smallest neighbor distance as compared to other distances studied while City Block had the highest distance. It has been stated in literature that Chebychev and Minkowski distances are concentric circle shape, this suggests the reason behind their equality of distance. It is therefore proposed that the data may be a concentric circle data.

Keyword-Genetic distances, Minkowski distance, Chebychev distance, Euclidean distance, City Block distance.

1. Introduction

Genetic distance is a measure of the genetic divergence between species or between populations within a species, whether the distance measures time from common ancestor or degree of differentiation. Populations with many similar alleles have small genetic distances.

Genetic distance is also defined as the term used to describe the number of differences or mutations between two cells of Y- chromosome DNA (Deoxyribonucleic acid) or mitochondrial DNA test results. A genetic distance of Zero means that there are no differences in the results being compared against one another, that is, there is an exact match. This is the meaning when comparing Y- chromosome DNA or mitochondrial DNA. For autosomal DNA comparisons, genetic distance relates to the size of a shared DNA segment. The genetic distance is then the length of the segment in centi Morgans. A centi Morgan is a unit of genetic distance that represents a 1% probability of recombination during meiosis. If two genes are 20 Centi Morgan (cM) apart; there is a 20% chance they will break apart during meiosis. The linkage distance or genetic distance is calculated by dividing the total number of recombinant gametes into the total number of gametes.

Some researchers have so far studied genetic distances and other genetic related studies, such as Onu, et al. (2021) studied the statistical bias in genetic model analysis with varying model parameters, where the relationships between heredity as response and the age and sex as predictors were considered. Avise & Aquadro (1982) summarized the multilocus allozyme literature on mean genetic distances (D's) between congeneric species and confamilial genera across the major vertebrate classes. Some salient trends emerged. Notably, mean D values among avian congeners were typically lower than those for other vertebrate groups. Congeneric species

of amphibians and reptiles often tended toward the high end of the mean genetic distance scale, whereas congeneric fishes and mammals generally were intermediate in magnitude of interspecific D's. Similar trends toward smaller genetic distances for birds than for other vertebrate groups also pertained just as seen in Glenn & John, 1998. Gentleman *et a*l. (2000) studied Distance Measures in DNA Microarray Data Analysis, using Minkowski based distance which include (Euclidean and City Block known as Manhattan distances) comparing them with correlation based distances like 1-pearsons distance, 1-spearman distance, etc. Other researchers which includes Lawan *et al.* (2001), Jessica *et al.* (2017), Daniel (1998), Steven (2002) and Glenn and John (1998) studied several genetic distances in different species, but none of the quoted literatures were able to compare the blood-group, gene frequencies characterizing four populations of four different countries (Eakimo, Bantu, English and Korea) using Minkowski, Euclidean, City Block and Chebychev distances in order to know the shape of the genetic data under study.

Genetic distance is useful for reconstructing the history of populations. For example, evidence from genetic distance suggests that Sub-Saharan African and Eurasian people diverged about 100,000 years ago. Genetic distance is also used for understanding the origin of biodiversity.

A gene is a unit of heredity which is transferred from a parent to offspring and is held to determine some characteristics of the offspring. It is also referred to as a distinct sequence of nucleotides forming part of a chromosome, the order of which determines the order of monomers in a polypeptide or nucleic acid molecule which a cell (or virus) may synthesize.

In biology, a gene is a basic unit of heredity and a sequence of nucleotides in DNA or RNA (Ribonucleic acid) that encodes the synthesis of a gene product, either RNA or protein. During gene expression, the DNA is first copied into RNA.

Various kinds of distance are significant in anthropological studies. We define genetic distance between two individuals (or between two populations) as the proportion of nonmatching nucleotide bases at homologous nucleotide sites between the genomes of two individuals (or of two populations). Since the matching nucleotide bases in these genomes are by and large the same by reason of descent from a common ancestor, genetic distance is also a measure of genetic OT phylogenetic divergence. Since there is no way to determine the nucleotide sequences of entire genomes, but there are data which can be used to deduce the nucleotide sequences of individual
genes and to sample the proportion of nonmatching nucleotide sites (i.e. the genetic distance) between genomes. Data for measuring genetic distance and hence for estimating genetic and phylogenetic divergence of genes, come from comparing the amino acid sequences of homologous polypeptide chains. Depending upon the proteins of polypeptide chains investigated, these sequences can help decipher genetic relationships among organisms at all taxonomic levels, from
individual differences within populations to differences between the animal and plant kingdoms. The most common event we acknowledge support by grants from the Systaltic Biology and U. S. - Japan Cooperative Science Programs of the National Science Foundation and Grant HDO 4815 from the Public Health Service. GOODMAN and LASKER in the ongoing process of evolution, the point mutation, often produces a difference in an amino acid residue in the sequences of proteins. The sequences are therefore an excellent method for estimating the distance created by establishment of point mutations in pedigrees and populations. Other types of biological tests

give results that can also be used for distance measurements. The sequence homology between DNA from two sources (HOYER et al. 1964, MARTIN and HOYER 1967, KOHNE 1970) can be determined by complementary base pairing. The correspondence of protein antigenic sites in different organisms can be evaluated by immunodiffusion comparisons in modified Ouchterlony plates (GOODMAN 1963a, 1963b, 1967, 1968; GOODMAN and MOORE 1971). This method works well in depicting genetic relatedness at the intermediate (generic through subordinal) taxonomic levels. Allelic frequency data, gathered by typing the polymorphic forms of enzymes and other proteins, usually by electrophoretic techniques, can measure in a rough way the genetic distances among individuals or populations at the lower (infrageneric) levels of species and within species. Typical blood cell isoantigens and serum protein allotypes by immunological procedures can permit the calculation of gene frequency differences in the same way.
Genetic distance is the divergence of genetic measurement between either species or populations within a species. (Yuan & Degui 2020). Experiments on Genome generate large and composite multivariate data sets. Machine learning approaches are important tools in microarray data analysis, for the purposes of identifying patterns in expression among genes and/or biological samples, and for predicting clinical or other outcomes using gene expression data. The ideal distance or similarity between the objects or things to be clustered or classified is inherent in Machine learning approach. Generally, any distance measure can be used with any machine learning algorithm. The choice of distance measure is probably more important than the choice of machine learning algorithm, and some attention should be paid to the selection of an appropriate measure for each problem. (Gentleman *et al.* 2000).

This ideal of distance is unambiguous in clustering procedures that operate directly on a matrix of pairwise distances between the objects to be clustered, for instance., partitioning around medoid (PAM) and hierarchical clustering (Kaufman and Rousseeuw, 1990). Certain supervised learning methods, such as nearest neighbor classifiers, also involve explicitly specifying a distance. Although the choice of distance may not be as transparent for other supervised approaches, observations are in fact assigned to classes on the basis of their distances from objects known to be in the classes. For instance, linear discriminant analysis is based on the Mahalanobis distance (Mardia *et al.* 1979, Ripley 1996a). The weighted gene voting scheme of Golub *et al*. (1999) is a variant of a special case of linear discriminant analysis, also known as naive Bayes classification. In addition, the distance and its behavior are intimately related to the scale on which measurements are made. The choice of a transformation and distance should thus be made jointly and in conjunction with the choice of a classifier or clustering procedure.

Analysis of genotypic data from neutral loci is an important method for describing the patterns of genetic variation within species and inferring the evolutionary processes that give rise to those patterns. Genotypic data are notoriously multivariate: the frequency of each allele at each locus is usually different in each population. Genetic distances are metrics that summarize these differences in an overall measure of differentiation for a pair of populations. Generally, a matrix of pair wise genetic distances between a set of populations is estimated. This matrix is then often visualized with phenograms, isolation by distance plots, principal component analysis, or multidimensional scaling plots. Many genetic distances have been developed, of which a few remain in regular use (Nei 1987 for a review of several genetic distances). Each of these genetic distances has unique evolutionary and statistical properties, and evolutionary relationships inferred from each genetic distances can be quite different. (Steven, 2002, Barker, 1999).

Glenn & John, (1998) stated that surveyed avian taxa on average, show significantly less genetic divergence than do same-rank taxa surveyed in other vertebrate groups, more notably are the amphibians and reptiles.

Steven, (2002), said that large sample sizes are warranted when populations are relatively genetically similar; and loci with more alleles produce better estimates of genetic distance.

Jessica *et al*. (2017) described that there was no significant correlation between pairwise genetic relatedness and multivariate trait distance among individuals.

Ruzzante, (1998) stated that The effect of number of alleles on sampling variance varied with the genetic measure considered.

Onu et al. (2021*) proposed* Grand Mean Absolute Deviation as a measure of the statistical bias that exists in the relationship between parents and the offspring in genetic studies.

At this point, the study will look at the various distances to be employed in this paper and their similarities and differences and they include:

City Block Distance:

The City block distance between two points, a and b, with k dimensions.

Chebyshev Distance:

This distance is also called maximum value distance. It studies the absolute magnitude of the differences between coordinates of a pair of points.

Euclidean Distance:

The Euclidean distance is the square root of a squared difference between a pair of points.

Minkowski Distance:

Minkowski distance is a generalized metric distance, its value is dependent on the shape of the object under study, which is determined by the value of λ.

2. Materials and Methods

City Block Distance:

In this study, the following distances are to be employed and compared appropriately.

The City block distance between two points, a and b, with k dimensions is given as:

$$\sum_{j=1}^{k} |a_j - b_j| \quad (1)$$

The City block distance is always greater than or equal to zero. The measurement would be zero for identical points and high for points that show little similarity.

Chebyshev Distance:

This distance is also called maximum value distance. It studies the absolute magnitude of the differences between coordinates of a pair of points. This distance measure can be used for both quantitative and ordinal data. It is given as

$$d_{ij} = max|a_{ik} - b_{jk}|$$

Euclidean Distance:

The Euclidean distance is the square root of a squared difference between a pair of points. It is given as

$$d_{ij} = \sqrt{(a_{ik} - b_{jk})^2} \quad (2)$$

Minkowski Distance:

Minkowski distance is a generalized metric distance, its value is dependent on the shape of the object under study, which is determined by the value of λ. For instance, if λ=1 it is concentric diamond and it becomes equal to the City Block distance, if λ=2 it is concentric circle and it becomes Euclidean distance if $\lambda > 3\ or\ = \infty$, it is concentric square and it becomes Chebychev Distance. It is given as

$$d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{n}|a_{ik} - b_{jk}|^{\lambda}} \quad (3)$$

We apply this data on these four distances and see how they behave and to know the shape of the data by knowing the particular distance that will be equal with the standard Minkowski distance. Here we cast our mind on the value of λ to the shape of the data.

3. Results and Discussion

The results of the analysis of this data with these distances were done using SPSS 23 and the results are as shown below:

Minkowski Distance Analysis

Table 1

Proximity Matrix

	Minkowski (2) Distance			
	Eakimo	Bantu	English	Korea
Eakimo	.000	1.295	1.044	.610
Bantu	1.295	.000	1.075	1.558
English	1.044	1.075	.000	1.116

Korea	.610	1.558	1.116	.000

This is a dissimilarity matrix

Euclidean Distance Analysis

Table 2

Proximity Matrix

	Euclidean Distance			
	Eakimo	Bantu	English	Korea
Eakimo	.000	1.295	1.044	.610
Bantu	1.295	.000	1.075	1.558
English	1.044	1.075	.000	1.116
Korea	.610	1.558	1.116	.000

This is a dissimilarity matrix

Squared Euclidean Distance Analysis

Table 3

Proximity Matrix

	Squared Euclidean Distance			
	Eakimo	Bantu	English	Korea
Eakimo	.000	1.677	1.089	.372
Bantu	1.677	.000	1.155	2.426
English	1.089	1.155	.000	1.245
Korea	.372	2.426	1.245	.000

This is a dissimilarity matrix

Chebychev Distance Analysis

Table 4

Proximity Matrix

	Chebychev Distance			
	Eakimo	Bantu	English	Korea
Eakimo	.000	.690	.607	.292
Bantu	.690	.000	.620	.935
English	.607	.620	.000	.574
Korea	.292	.935	.574	.000

This is a dissimilarity matrix

City Block Distance Analysis

Table 5

Proximity Matrix

	City Block Distance			
	Eakimo	Bantu	English	Korea
Eakimo	.000	3.791	3.188	1.973
Bantu	3.791	.000	3.027	4.005
English	3.188	3.027	.000	3.288

Korea	1.973	4.005	3.288	.000

This is a dissimilarity matrix

Discussion of Results

The study of various distances for blood-group gene frequencies characterizing four populations, reveals that Chebychev distance has the lowest distance between a pair of countries, followed by the Minkowski distance which is accurately equal to the Euclidean distance and then the City Block distance. Bantu and Korea has the highest neighbor distance followed by Bantu and Eakimo, while the smallest distance is between Korea and Eakimo. Chebychev distance proved to be the best distance for this study since it had the smallest neighbor distance as compared to other distances studied. This is because, the smaller the distance between a pair of points, the similar they are. Whereas Euclidean distance shows that the data so far studied has a shape of concentric circle, this was revealed because of the equality of Minkowski distance and the Euclidean distance. For the Minkowski and Euclidean distances, it was found that the correlation between Eakimo and Korea has the smallest distance while Korea and Bantu has the highest distance. For Squared Euclidean distance, it was observed that Eakimo and Korea has the highest distance, while English and Eakimo has the smallest distance. For the Chebychev distance, Eakimo and Korea has the smallest distance, while Bantu and Korea has the highest distance. Finally, for the City block distance, Bantu and Korea has the highest distance while Eakimo and Korea has the lowest distance.

4. Conclusion

This study concludes that the data used in this analysis is concentric circle, underscoring why the Minkowski and Euclidean distances of such data were equal. Also, that correlation study of these countries for each of these distances showed that Chebychev distance has the smallest distance among all the distances study followed by the duo of Euclidean and Minkwoski distances.

Recommendations

This study recommends to statisticians and other related disciplines for the study of the effect of blood-group gene frequencies on different countries that;

Minkowski and Euclidean distances are better distance formulas to be used when the data is suspected to be a concentric circle.

The blood-group gene frequencies for one country varies with another country.

The countries Eakimo and Korea has the smallest distance or are more similar than others for Minkowski and Euclidean distances, for squared Euclidean distance, the English and Eakimo has the smallest distance, for Chebychev distance the Eakimo and Korea has the smallest, while for City block distance the Eakimo and Korea has the smallest distance for the blood-group gene frequencies.

References

Avise, J. C. (1983). Commentary. Pp. 262–270 *in* A. H. BRUSH and G. A. CLARK, eds. Perspectives in ornithology. Cambridge University Press, Cambridge. *Molecular markers, natural history and evolution. Chapman and Hall.*

Avise, J. C., and C. F. Aquadro. (1982). A comparative summary of genetic distances in the vertebrates. *Evolution. Biological journal.* ***15***: 151–184.

Barker, J. S. F., (1999). A Global protocol for determining genetic distances among domestic livestock breed.

Cavalli-Sforza, L. L., & Edwards, A. W. F., (1983). Phylogenetic Analysis Models and Estimation Procedures. *International Laboratory of Genetics and Biophysics, Naples, and Pavia Section, Istituto di Genetica, Universita di Pavia.*

Glenn, C. J. & John, C. A., (1998). *A Comparative summary of Genetic Distances in Vertebrates from the mitochondrial cytochrome b Gene.*

Jessica, M. A., Katherine, D., Richard, K. G., Susan, L. W., & John, J. S. (2017). Genetic Distance Predict traits differentiation at the subpopulation but not the individual level eelgrass Zostera marina. *Wily, ecology and evolution.*

Laval, G. Magali, S., & Chevalet, C., (2001). Measuring genetic distances between breeds: use of some distances in various short term evolution models. *Genet. Sel. E, 34, 481–507 481 INRA, EDP Sciences.*

Lydeard, C., and K. J. Roe. (1997). The phylogenetic utility of the mitochondrial cytochrome *b* gene for inferring relationships among Actinopterygian fishes. Pp. 285–303 *in* T. D.

Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, **89**, 583Ð590.

Nei M (1987) Molecular Evolutionary Genetics. Columbia University Press, New York.

Onu, O. H., George, D. S., Uzoamaka, C. E., & Okerengwu, B. (2021). The Statistical Bias in Genetic Model analysis with varying Model parameters. *International Journal of Research (IJR), 8*(6), 154-166.

Ruzzante, D., (1998). A Comparison of several methods of genetic distances and population structure with microsatellite data: Bias and sampling variance. *Researchgate, 55, 1-14.*

Steven, T. K., (2002). Evolutionary and statistical properties of three genetic distances, *Molecular Ecology, 11, 1263–1273*

Chart 1 : Data for the study Sourced from: L. L. CAVALLI-SFORZA AND A. W. F. EDWARDS (1983)

Allele	Eakimo	Bantu	English	Korea
A1	.29	.10	.21	.22
A2	.00	.09	.07	.00
B	.03	.12	.06	.21
0	.68	.69	.07	.57
CDE	.00	.00	.00	.01
CDe	.50	.14	.42	.62
Cde	.49	.10	.14	.31
CDe	.11	.60	.03	.06
Cde	.00	.02	.10	.00
cdE	.00	.00	.01	.00
Cde	.00	.23	.39	.00
MS	.17	.09	.24	.02
Ms	.67	.48	.30	.46
NS	.00	.04	.07	.06
FYa	.16	.39	.39	.45
FYb	.75	.06	.42	1.00
Dia	.25	.94	.58	.01
Dib	.00	.00	.00	.03