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 thatChebychevand Minkowski distances are concentric circle shape, this suggests the reason behind theirequality 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 comparedagainst 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 centiMorgans. A centiMorgan is a unit of genetic distance that represents a 1% probability of recombination during meiosis. If two genes are 20 CentiMorgan (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 various distances especially in Genetics and other genetic related studies, such as (Rizwan et al 2009, Raneem et al 2020 and Black 2006), but 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 al (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 Jessica *et al*. (2017), 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 similarities of these countries in terms of their blood-group, gene frequenciesand also, to know the shape of the genetic data under study. 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). The sequence homology between DNA from two sources as stated by Hoyer 1967,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. This method works well in depicting genetic relatedness at the intermediate (generic through subordinal) taxonomic levels. Goodman & Lasker (1964) stated that 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.

In the comparison of distances, Rizwan et al. (2009) observed that minkowski coefficient best approximates road distance by 1.54, while 1.31 best approximates travel time. It was also found to be a good predictor of road distance which then provides the best single model for traveling patients from patient's residence to the hospital. While the Euclidean metric and minkowski metric are alternatively used for regression model. The study stated that minkowski method gave more reliable results than the Euclidean distance. Euclidean distance under estimates road distance and city block also known as Manhattan distance over estimates road distance. It is the minkowski distance that overcomes these disagreements.

According to Raneemetal (2020), minkowski distance is a generalizer of other distances, such as the manhattan (city block) and Euclidean distances. The result stated that data sets have low dimensions, givesthe best average results for different distance, (Euclidean, manhattan or city block/ Euclidean, chebyshev and chebychev respectively). Manhattan distance measure was recommended for high dimensional data as it shows the highest average values of purity for the remaining high dimensional data set. Chebyshev was found to be the worst clustering results.Also see (Anton, 2013).

Genetic distance isthe divergence of genetic measurement between either species or populations within a species. (Yuan &Degui 2020). Also, Goodman & Lasker (1964) defined Genetic distance between two populations or two individuals as the proportion of nonmatching nucleotide bases at homologous sites between the genomes of the two populations or individuals.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.

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).

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 pairof 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 $\lambda$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

Thestudy 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.The result of the high similarity between Minkowski and Euclidean distances affirms the statement of Raneem (2020), which states that Minkowski distance as a generalizer of other distances and it also affirms that minkowski coefficient best approximates road distance as stated by Rizwan et al (2009). 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 Bantuand 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;

1. Minkowski and Euclidean distances are best distance formulas to be used when the data is suspected to be a concentric circle.
2. The blood-group gene frequencies for one country varies with another country.
3. 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

Anton, H.(2013). Elementary linear Algebra, Binder ready version. John wiley & sons.

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.

Black, P.E., (2006). Manhaltan Distance'dictionary of algorithms and data structures. http://xlinux.nist.gov/dads//.

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.*

Gentleman, R., Ding, B., Dudoit, S. & Ibrahim, J. (2000). Distance measures in DNA microarray data analysis. Bioinformatics and computational Biology, 189-208.

Goodman, M. & Lasker, G.W. (1964). Measurement of distance and propinquity in Anthropological studies. *Springer*, 5-21.

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.*

Kaufman L.P. & Rousseeuw P.J. (1990). Finding groups in data: An introduction to cluster analysis. Wiley series in probability and statistics.

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.*

Mardia, K.V., Kent, J.T & Bibby, J.M. (1979). Multivariate Analysis. Academic press, probability and statistics.

Martin, M.A., Hoyer, B.H. (1967). Adenine plus thymine and guanine plus cytosine enriched fractions of animal DNA's as indicators of polynuclertide homologies. Journal of molecular Biology, 27(1), 113-129.

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.

Raneem, Q., Hossam, F., Ibrahim, A., Merelo, J.J & Pedro, A.C. (2020). Emperical Evaluation of distance measures for nearest point with indexing ratio clustering algorithm. In proceedings of the12th international joint conference on computational intelligence.

Rizwan, S., Stefania, B., Merril, L.K & William, A.G. (2009). Comparison of distance measures in spatial analytical modeling for health service planning. BMC Health services research, 9(200).

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.*

Shraddha, P& Suchita, G. (2011). Acomparative study on distance measuring approaches for clustering. International journal of research in computer science, 2(1), 29-31.

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

Yuan, W & Degui, Z. (2020). Genealogical Search using whole-genome genotype profiles. Responsible Genomic Data Sharing.