Analysis of Genetic Relationship Among 11 Iranian Ethnic Groups with Bayesian Multidimensional Scaling Using HLA Class II Data

Najaf Zare[1], Shirin Farjadian[2] and Samaneh Maleknia[1,*]

[1]*Department of Biostatistics, Shiraz University of Medical Sciences, Shiraz, Iran*

[2]*Department of Immunology, Shiraz University of Medical Sciences, Shiraz, Iran*

Abstract: *Background*: The key feature of Bayesian methods is their lack of dependence on defaults necessary for classical statistics. Because of the high volume of simulation, Bayesian methods have a high degree of accuracy. They are efficient in data mining and analyzing large volumes of data, and can be upgraded by entering new data.

Objective: We used Bayesian multidimensional scaling (MDS) to analyze the genetic relationships among 11 Iranian ethnic groups based on HLA class II data.

Method: Allele frequencies of three HLA loci from 816 unrelated individuals belonging to 11 Iranian ethnic groups were analyzed by Bayesian MDS using R and WinBUGS software.

Results: like the results of correspondence analysis as a prototype of classical MDS analysis, the results of Bayesian MDS also showed Arabs from Famur, Balochis, Zoroastrians and Jews to be separate from other Iranian ethnic groups. Decreases stress in Bayesian MDS method compared to classical method revealed the accuracy of Bayesian MDS for HLA data analyses.

Conclusion: This study reports the first application of Bayesian multidimensional scaling to HLA data analysis with Nei's D_A genetic distances. Stress reduction in Bayesian MDS compared to classical MDS showed that the Bayesian approach can improve the accuracy of genetic data analysis.

Keywords: Bayesian methods, Multidimensional scaling, Anthropological study, Immunogenetics, R and WinBUGS software.

INTRODUCTION

Bayesian multidimensional scaling (MDS) is one of the graphical multivariate analyses which is often used for genetic data analysis [1-3]. Bayesian theory which first proposed by Thomas Bayes in eighteenth century is based on formulating probability distributions to express uncertainty about unknown quantities [4, 5], Bayesian statistics has been widely applied in different fields from economical sciences to medical researches and genetic studies [6-9]. The most crucial characteristic of MDS is its simplification of complex data analysis by reducing the dimensions. The key feature of Bayesian methods is their lack of dependence on defaults necessary for classical statistics [5].

Human leukocyte antigen (HLA) genes encode the highly polymorphic molecules responsible for antigen presentation to T lymphocytes. Because of their high variability, HLA molecules are also considered the main problem in transplantation [10]. The frequency of HLA alleles differs widely among populations [11] and genetic distances among populations based on HLA data can be helpful in choosing better donor candidates for transplantation [12].

Correspondence analysis or classical MDS are usually used to explore genetic relationships among populations. Since Bayesian approach to MDS offers some advantages over similar classical MDS procedures, in this study we used Bayesian MDS to investigate if this method can improve our genetic analysis compared to classical MDS which previously has been used for analysis of these data.

METHODS

Genetic Data

Previously published HLA data (DQA1, DQB1 and DRB1 allele frequencies) from 816 unrelated healthy individuals belonging to 11 Iranian ethnic groups: Pars (72 individuals), Zoroastrian (65 individuals), Baloch (100), Arab (50 individuals from Ahvaz and 84 from Famur), Jew (91 individuals), Kurd (100 individuals), Azeri (100 individuals), Bakhtiari (50 individuals) and Lur (50 individuals from Luristan and 54 from Yasouj) was used in this study [13]. The data were analyzed by Bayesian MDS with Nei's D_A genetic distance [14]. All analyses were done with R (http://www.r-project.org) and WinBUGS (http://www.mrc-bsu.cam.ac.uk/bugs) software.

*Address correspondence to this author at the Department of Biostatistics, Shiraz University of Medical Sciences, Zand St., 71348-45794 Shiraz, Iran; Tel: +98 9351304346; E-mail: maleknias@gmail.com

E-ISSN: 1929-6029/13

Statistical Method

Because of its ability to determine relationships among datasets and due to its special type of geometric representation, MDS is widely used in different branches of science. A specific algorithm for a set of proximities is used to select a particular type of spatial representation, and then modeling is performed [15]. When either the vectors of observations or the distances between stimuli are available, the best type of geometrical representation is classical MDS, also known as metric MDS. Graphical displays are obtained from a set of transformations on the D matrix using eigen values and eigen vectors to create a new matrix, $\hat{D}$ [15-17].

Different types of distance values have previously been used in MDS analyses [16]. In this study, Nei's genetic distance, D_A, was calculated with the following [18].

$$D_A = \frac{1}{r}\sum_{j=1}^{r}(1-\sum_{i=1}^{m_j}\sqrt{x_{ij}y_{ij}})$$

where x_{ij} and y_{ij} were the frequencies of the *i*th allele at the *j*th locus in populations X and Y, m_j was the number of alleles at the *j*th locus, and r was the number of loci studied.

Goodness-of-fit tests were used to ensure that the number of dimensions was appropriate. In this study the stress value was used to evaluate the goodness-of-fit. The stress value was obtained using arrays of D and $\hat{D}$ matrices with the following formula [2, 16, 17].

$$Stress = \sqrt{\frac{\sum_{i=1}^{n}\sum_{j=1}^{n}(d_{ij}-d_{ij}^{(0)})^2}{\sum_{i=1}^{n}\sum_{j=1}^{n}(d_{ij}^{(0)})^2}}$$

Based on the criteria of Kruskal and Wish; stress <0.025 was considered excellent, between 0.025 and <0.05 was considered good, between 0.05 and <0.1 was considered fair, and stress ≥0.2 was considered poor [19].

Bayes' theorem for parameter estimation was obtained by calculating the posterior distribution of θ based on the given matrix D:

$$p(\theta / D) = \frac{p(D/\theta)p(\theta)}{p(D)} = \frac{likelihood \times prior}{\text{int}\,egrated\ likelihood}$$

$$p(D) = \int_{all\ values\ of\ \theta} p(D/\theta)p(\theta)d(\theta)$$

Where *p*(D) is the integrated likelihood [20]. The integral was calculated by Markov chain Monte Carlo simulation with the Metropolis–Hastings algorithm [21]. Heidelberger and Welch criteria were used to determine the diagnostic convergence of the chain [22].

The posterior density functions of the unknown parameters (X, σ^2, **^)** were calculated as follows:

$$\pi(X,\sigma^2,\Lambda,p \mid D) \propto (2\pi)^{-\frac{m}{2}}\sigma^{-m}$$
$$\times\exp\left[-\frac{1}{2\sigma^2}SSR-\sum_{i>j}\log\Phi\left(\frac{\delta_{ij}}{\sigma}\right)\right]\times(2\pi)^{-\frac{np}{2}}\prod_{j=1}^{p}\lambda_j^{-(n/2)}\exp\left[-\sum_{j=1}^{p}\frac{1}{2\lambda_j}s_j\right]$$
$$\times\Gamma(a)^{-1}b^{a}(\sigma^2)^{-(a+1)}\exp\left[-\frac{b}{\sigma^2}\right]\times\Gamma(a)^{-p}\prod_{j=1}^{p}\beta_j^{\alpha}\lambda_j^{-(\alpha+1)}\exp\left[-\frac{\beta_j}{\lambda_j}\right]$$

where *p* was the number of dimensions and *D* was the matrix of observed distances.

The software used in this study, WinBUGS and R, used classical MDS results for the parameters of prior distributions and then estimated the posterior distributions and values with Markov chain Monte Carlo simulation.

The appropriate number of dimensions was obtained with the Multidimensional Scaling Information Criterion (MDSIC) according to the formula:

Table 1: Measurement of MDSIC in Any Data Source for Different Number of Dimension (Minimum MDSIC in each Locus is Shown in Bold)

Data source	One dimension	Two dimensions	Three dimensions	Four dimensions	Optimal number of dimensions
DQA	263.246	**261.034**	287.844	331.292	2
DQB	-122.122	-189.131	**-196.993**	-167.535	3
DRB	-114.533	-167.114	-219.342	-243.574	4
All	-202.859	-270.912	**-318.694**	**-299.423**	3

Table 2: The Comparison of Stress in Classical MDS and Bayesian MDS in Optimal Dimensions (Three Dimensions)

Data source	Minimum MDSIC	Stress for CMDS	Stress for BMDS	Improvement rate
DQA	261.0340	0.31466	0.19705	0.11761
DQB	-196.9927	0.17784	0.11065	0.06719
DRB	-243.5740	0.22851	0.05697	0.17154
All	-318.6939	0.16019	0.07567	0.08452

CMDS, classic multidimensional scaling; BMDS, Bayesian multidimensional scaling.

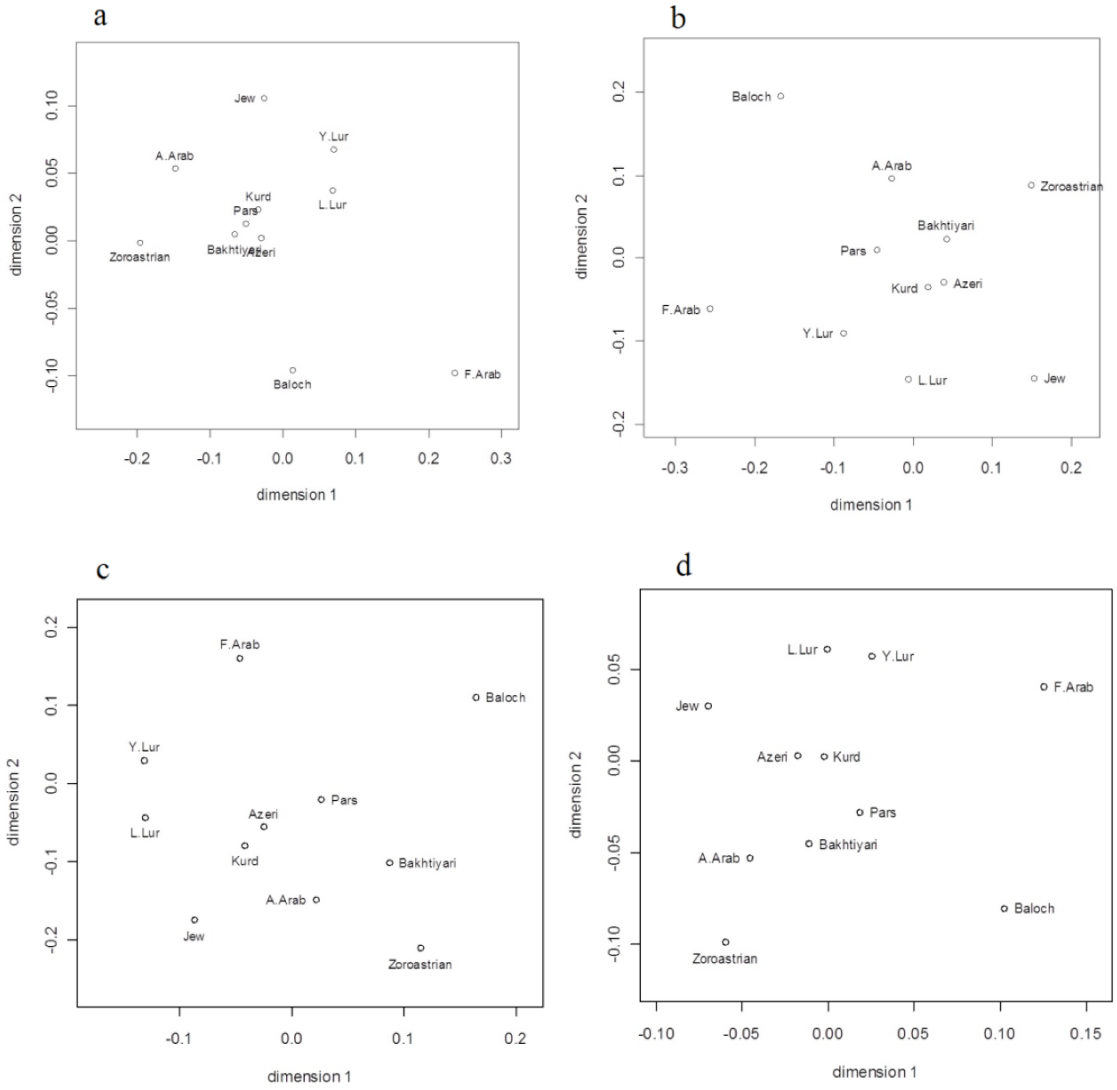

Figure 1: Bayesian multidimensional scaling analysis showing bi-dimensional representation of the genetic relationship among 11 Iranian ethnic groups based on allele frequencies on HLA-DQA1 (**a**), DQB1 (**b**), DRB1 (**c**), and all three loci (**d**).

$$MDSIC_1 = (m-2)\log SSR_1 \rightarrow SSR_1 = \sum_{i\ j} (d_{ij} - \delta_{ij})^2$$

.

.

$$MDSIC_p = MDSIC_1 + \sum_{j=1}^{p-1} LR_j$$

The optimal number of dimensions was obtained by minimizing MDSICp [2].

RESULTS AND DISCUSSION

Bayesian MDS was used to analyze the genetic relationships among 11 Iranian ethnic groups based on HLA class II data. Table **1** shows the minimum MDSIC

detected in different numbers of dimensions for the different HLA loci. The improvements in stress with Bayesian MDS compared to classical MDS are shown in Table **2**. Although the optimum dimensions were different for each locus, two-dimensional representation was used for further analysis.

The genetic relationships among 11 Iranian ethnic groups using Bayesian MDS analysis are depicted in Figure **1**. The distribution pattern of the ethnic groups differed somewhat depending on which locus was used as the source of genetic information. When the allele frequencies at all three loci were considered, Azeris, Kurds, Parsees, Bakhtiaris and Arabs from Ahvaz clustered together, and Lurs from Luristan and Yasouj were located in a separate cluster. Arabs from Famur, Balochis, Zoroastrians and Jews separated from other ethnic groups as outliers. This might be explained by religious or cultural differences between these groups and other ethnic groups studied here.

Similar results were reported previously based on correspondence analysis [13]. However that earlier study found that Arabs from Famur, Balochis, Zoroastrians and Jewes were well separated from other ethnic groups and were outliers, whereas the remaining nine ethnic groups were located in a single cluster.

To our knowledge, this is the first study to implement Bayesian MDS for HLA data analysis using D_A genetic distances. Like MDS, correspondence analysis is a multivariate data reduction technique with a graphical output which uses the raw data to create a two-dimensional matrix [23]. For HLA data analyses, MDS with D_A and a one-dimensional matrix are generally recommended. We calculated D_A based on distances among ethnic groups, and constructed a one-way matrix but MDS analyzed with Bayesian methods. As previous articles [24], the decreases stress in Bayesian MDS compared to classical MDS showed that the accuracy of MDS can be improved with Bayesian techniques. As shown by our concurrent use of all three alleles to calculate D_A, highly polymorphic genes or the simultaneous study of different genetic loci are potentially helpful in enhancing the accuracy of estimates of genetic proximity with this approach.

ACKNOWLEDGEMENTS

This article is based on research done in partial fulfillment of the requirements for the MSc degree in biostatistics awarded to Samaneh Maleknia by Shiraz University of Medical Sciences. We thank Prof. K. Okada, University of Tokyo, for his valuable help in running WinBUGS from R for simulations, and K. Shashok (Author AID in the Eastern Mediterranean) for improving the use of English in the manuscript.

SUPPORT

The research was financially supported by Shiraz University of Medical Sciences (Grant No. 88-4689).

REFERENCES

[1] DeSarbo WS, Kim Y, Fong D. A Bayesian multidimensional scaling procedure for the spatial analysis of revealed choice data. J Economet 1998; 89(1-2): 79-108. http://dx.doi.org/10.1016/S0304-4076(98)00056-6

[2] Oh MS, Raftery AE. Bayesian multidimensional scaling and choice of dimension. J Am Statist Assoc 2001; 96(455): 1031-44. http://dx.doi.org/10.1198/016214501753208690

[3] Park J, DeSarbo WS, Liechty J. A hierarchical Bayesian multidimensional scaling methodology for accommodating both structural and preference heterogeneity. Psychometrika 2008; 73(3): 451-72. http://dx.doi.org/10.1007/s11336-008-9064-1

[4] Stigler SM. Who discovered Bayes's theorem? Am Statist 1983; 37(part 4a): 290-6.

[5] Lindley DV, Lindley D. Bayesian statistics: A review: SIAM 1972; pp. 1-9. http://dx.doi.org/10.1137/1.9781611970654.ch1

[6] Kim C-J, Nelson CR. Has the US economy become more stable? A Bayesian approach based on a Markov-switching model of the business cycle. Rev Econom Statist 1999; 81(4): 608-16. http://dx.doi.org/10.1162/003465399558472

[7] Corander J, Waldmann P, Sillanpää MJ. Bayesian analysis of genetic differentiation between populations. Genetics 2003; 163(1): 367.

[8] Boys RJ, Henderson DA. A Bayesian approach to DNA sequence segmentation. Biometrics 2004; 60(3): 573-81. http://dx.doi.org/10.1111/j.0006-341X.2004.00206.x

[9] Ashby D. Bayesian statistics in medicine: a 25 year review. Statist Med 2006; 25(21): 3589-631. http://dx.doi.org/10.1002/sim.2672

[10] Rogers NJ, Lechler RI. Allorecognition. Am J Transplant 2001; 1(2): 97-102. http://dx.doi.org/10.1034/j.1600-6143.2001.10201.x

[11] Arnaiz-Villena A, Iliakis P, González-Hevilla M, Longas J, Gómez-Casado E, Sfyridaki K, *et al*. The origin of Cretan populations as determined by characterization of HLA alleles. Tissue Antigens 1999; 53(3): 213-26. http://dx.doi.org/10.1034/j.1399-0039.1999.530301.x

[12] Zachary AA, Kopchaliiska D, Jackson AM, Leffell MS. Immunogenetics and immunology in transplantation. Immunol Res 2010; 47(1-3): 232-9. http://dx.doi.org/10.1007/s12026-009-8154-1

[13] Farjadian S, Ota M, Inoko H, Ghaderi A. The genetic relationship among Iranian ethnic groups: an anthropological view based on HLA class II gene polymorphism. Mol Biol Rep 2009; 36(7): 1943-50. http://dx.doi.org/10.1007/s11033-008-9403-4

[14] Nei M. Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 1973; 70(12): 3321-3. http://dx.doi.org/10.1073/pnas.70.12.3321

[15] Jobson JD. Applied multivariate data analysis: Categorical and Multivariate Method. 4th ed: Springer 1998; pp. 760-764.

[16] Cox TF, Cox MAA. Multidimensional scaling. 2nd ed: CRC Press 2001; chapter 1-2.

[17] Rencher AC. Methods of multivariate analysis. 2nd ed: Wiley-Interscience; 2002; chapter: 15.2; pp. 504-507.

[18] Nei M, Tajima F, Tateno Y. Accuracy of estimated phylogenetic trees from molecular data. J Mol Evolut 1983; 19(2): 153-70. http://dx.doi.org/10.1007/BF02300753

[19] Kruskal JB. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 1964; 29(1): 1-27.

[20] Smith AF, Gelfand AE. Bayesian statistics without tears: a sampling–resampling perspective. Am Statist 1992; 46(2); 84-88.

[21] Andrieu C, Doucet A, Holenstein R. Particle markov chain monte carlo methods. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 2010; 72(3): 269-342. http://dx.doi.org/10.1111/j.1467-9868.2009.00736.x

[22] Cowles MK, Carlin BP. Markov chain Monte Carlo convergence diagnostics: a comparative review. J Am Statist Assoc 1996; 883-904. http://dx.doi.org/10.1080/01621459.1996.10476956

[23] Rencher AC. Methods of multivariate analysis. 2nd ed: Wiley-Interscience; 2002; chapter: 15.3; pp. 514-530.

[24] Okada K, Shigemasu K. BMDS: A Collection of R Functions for Bayesian Multidimensional Scaling. Appl Psychol Measurem 2009; 33(7): 2. http://dx.doi.org/10.1177/0146621608321761

Received on 01-04-2013 Accepted on 19-07-2013 Published on 31-07-2013

http://dx.doi.org/10.6000/1929-6029.2013.02.03.5