A Graphical Approach for Representing Disequilibrium in Population Genetic Study

Mira Park
Eulji University School of Medicine
143-5 Yongdu-dong Choong-ku
Taejon, Korea
mira@emc.eulji.ac.kr

Jae Won Lee
Department of Statistics
Korea University
5-1 Anam-dong Seongbuk-ku
Seoul. Korea
jael@mail.korea.ac.kr

1. Introduction

To explore the relationships among frequencies for sets of alleles, within or between loci, is one of the first analyses in population genetic study. The general question is whether the frequency of a set of alleles is the same as the product of each of the separate allele frequencies. For two alleles of a single locus, Hardy-Weinberg disequilibrium is tested and for an allele from each of two loci, linkage disequilibrium is tested. The traditional goodness-of-fit test and likelihood ratio test have been used to test association, and for smaller samples, exact tests are used (Emigh, 1980; Guo and Thomson,1992). However, it is more useful if we can quantify and graphically represent this information. In this study, we suggest graphical methods to find associations between alleles. We also analyze the STR data of Korean population as an illustration.

2. Quantification and Graphical Methods

There are various multivariate techniques to present relationships between variables by graph. In this study, we consider the correspondence analysis for single locus situation. When we have r alleles, we can construct rxr matrix $F=(f_{ij})$, where f_{ij} is homozygote counts of allele i divided by total counts n and $f_j=f_{ji}$ is heterozygote counts of pair with allele (i, j) divided by 2. Thus F is a symmetric matrix.

Classical correspondence analysis of rxc matrix can be performed by singular value decomposition of

$$G=Dr^{-1/2}(F-rc')\,Dc^{-1/2}=UD_\lambda V'$$

where $r=(f_{1+}, \,..f_{r+})'$, $c=(f_{+1}, f_{+c})'$, $Dr=diag(f_{1+}, \,..f_{r+})$, $Dc= diag(f_{+1}, .., f_{+c})$, $U'U=V'V=I$ and D_λ =is kxk diagonal matrix with elements ${}_1\lambda$ … ${}_k\lambda$. And the principal co-ordinates of the row and column profiles are given by

$$X=Dr^{-1/2}UD_{\lambda} \text{ and } Y= Dc^{-1/2}VD_{\lambda}.$$

(Greenacre and Hastie, 1987). For genetic data, row and column profiles are the same because the correspondence matrix is symmetric. And it is enough to plot the row (column) profiles solely. We can show least square approximation of the row(column) $\div^2$ distance and interpret the distance between the points. If we plot

$$X^*=UD_{\lambda}^{\acute{a}} \text{ and } Y^*= VD_{\lambda}^{1-\acute{a}}$$

instead of X and Y, we can obtain the biplot of correspondence matrix. Since the element of G is

$$\div_{ij} = (f_{ij} - f_{i+}f_{+j}/f_{++}) \;/\quad f_{i+} \quad f_{+j}$$

whose square is proportional to the contribution of the (i,j)th cell to $\div^2$ statistic for independence of the row and column. By taking $\acute{a}$=1/2 and plot X^* (or Y^*), we can get inner-product approximation to the individual terms $\div_{ij}$ of $\div^2$ statistic. And this plot represents relationships among frequencies for sets of alleles within locus and which alleles contribute to break H-W equilibrium. Similarly, multiple correspondence analysis is used to explore the relationships among allele frequencies between loci.

3. Concluding Remarks

We analyze the 17 STR data in Korean population. From graphs by simple and multiple correspondence analysis, we can detect the associations between alleles and check equilibrium. These results are compared to the graphs from canonical correlation biplot proposed by Park and Huh(1996).

REFERENCE

Emigh TH (1980) A comparison of tests for Hardy-Weinberg law. Biometrics 36 627-642.

Guo SW and Thomson EA (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles, Biometrics, 48, 361-372.

Park MR and Huh MH (1996) Canonical corrrelation biplot. The Korea Communications in Statistics 3(1): 11-19

Greenacre M and Hastie T (1987) Geometric interpretation of correspondence analysis. Journal of the American Statistical Association 82: 437-337.

RESUME

We suggest quantification and graphical methods to represent relationships among allele frequencies within or between loci. Algorithms for correspondence and multiple correspondence analysis are used for modified contingency tables. STR data of Korean population are analyzed as an illustration.