Available online at www.sciencedirect.com

ScienceDirect

Computational Statistics & Data Analysis 51 (2007) 3145 – 3155

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Correspondence analysis approach for finding allele associations in population genetic study

Mira Park[a], Jae Won Lee[b,*], Choongrak Kim[c]

[a]*Department of Pre-medicine, Eulji University, 143-5 Yongdu-dong, Chung-ku, Daejeon 301-832, Republic of Korea*
[b]*Department of Statistics, Korea University, Anam-dong, Sungbuk-ku, Seoul 136-701, Republic of Korea*
[c]*Department of Statistics, Pusan National University, Jangjeon-dong, Geumjeong-ku, Pusan 609-735, Republic of Korea*

Received 14 April 2006; received in revised form 26 July 2006; accepted 8 September 2006
Available online 2 October 2006

Abstract

In population genetic study, one of the first analyses is to explore the relationships among the frequencies of alleles within or between loci. Hardy–Weinberg equilibrium is tested for two alleles of a single locus, and the linkage disequilibrium is tested for an allele from each of two loci. Although the equilibrium plays an important role and often serves as a basis for genetic inference, research on the graphical representation of this information is rare. In this study, we consider correspondence analysis and biplot approaches as tools for finding associations between alleles. We also propose the supplementary data method to compare allele frequencies of several populations from different studies. These approaches provide the graphical representation which makes it easy to interpret the patterns of disequilibrium and to compare the allele frequencies between populations. These proposed methods are illustrated with numerical examples.
© 2006 Elsevier B.V. All rights reserved.

Keywords: Correspondence analysis; Biplot; Hardy–Weinberg equilibrium; Linkage disequilibrium; Supplementary data analysis

1. Introduction

Once the allele frequencies and the genotype frequencies have been estimated, one of the first analyses of population genetic data is to look for associations among the frequencies of alleles within or between loci. When there are no disturbing forces such as selection, mutation or migration that would change the allele frequencies over time, and when there is a random mating in very large populations, these pairs of alleles are known not to be associated. A consequence of this independence is that the genotype frequencies are equal to the product of allele frequencies (Weir, 1996). The differences between joint frequencies and the products of individual frequencies are called disequilibrium coefficients. For two alleles at a single locus, we test whether the Hardy–Weinberg disequilibrium is zero, and for an allele from each of two loci, we test whether the linkage disequilibrium is zero. Furthermore, it often happens we want to compare the allele frequencies of several populations.

Despite its importance in the field of population genetics and its role as a basis for the genetic inference, only the p-values of the tests are reported in most researches (cf. Iwasa et al., 1997; Budowle et al., 1997; Gehrig et al., 1999;

[*] Corresponding author. Tel.: +82 2 3290 2237; fax: +82 2 924 9895.
E-mail addresses: mira@eulji.ac.kr (M. Park), jael@korea.ac.kr (J.W. Lee), crkim@pusan.ac.kr (C. Kim).

0167-9473/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2006.09.002

Fregeau et al., 1998). A graphical representation of the association between alleles is helpful, especially when there are multiple alleles, to understand the properties of the given data, but there are only a few researches for graphical approach to analyze the allele frequency data (Sjerps et al., 1995; Abecasis and Cookson, 2000). In this study, we consider the correspondence analysis (CA) and biplot approaches as the graphical methods for showing the allele relationships within and between loci. These plots can check whether the Hardy–Weinberg equilibrium or the linkage equilibrium is satisfied or not. We also suggest a supplementary data method to compare several populations from different study. In Section 2, we describe the graphical methods to represent the association between alleles in a single locus. We extend these methods to the two loci cases in Section 3. In Section 4, we propose a supplementary data method for comparing several populations from different studies. In each section, we provide the examples using real and/or simulated STR data.

2. Graphical approach for finding allele association within a locus

Consider a locus which has k alleles $A_1, A_2, \ldots, A_k$. In Hardy–Weinberg equilibrium, the frequency of a homozygote A_iA_i becomes p_i^2 and that of a heterozygote A_iA_j becomes $2p_ip_j$, where p_i is the population frequencies of allele A_i. For allele A_i and A_j, the disequilibrium coefficients is defined as $D_{ij} = p_{ij} - p_ip_j$, where p_{ij} is the population frequencies of allele A_{ij}. If a sample with size n is from a population of interest and x_{ij}^* $(1 \leqslant j \leqslant i \leqslant k)$ is the observed count of genotype A_iA_j, then under Hardy–Weinberg equilibrium,

$$\chi^2 = \sum_i \frac{(x_{ii}^* - n\hat{p}_i^2)^2}{n\hat{p}_i^2} + \sum_i \sum_{j \neq i} \frac{(x_{ij}^* - 2n\hat{p}_i\hat{p}_j)^2}{2n\hat{p}_i\hat{p}_j} \tag{1}$$

follows χ^2 distribution with $k(k-1)/2$ degrees of freedom, where $\hat{p}_i = x_{ii}/n + \frac{1}{2}\sum_{j\neq i} x_{ij}$ (Weir, 1996).

Let $X = (x_{ij})$ be $k \times k$ symmetric matrix whose diagonal elements are $x_{ii} = x_{ii}^*$ and the non-diagonal elements are $x_{ij} = x_{ji} = x_{ij}^*/2$. Define G and its singular value decomposition as

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

where $F=(f_{ij})$, $f_{ij}=x_{ij}/n$, $r=(f_{1+}, \ldots, f_{r+})'$, $c=(f_{+1}, \ldots, f_{+c})'$, $f_{i+}=\sum_j f_{ij}$, $f_{+j}=\sum_i f_{ij}$, $D_r = diag(f_{1+}, \ldots, f_{r+})$ and $D_c = diag(f_{+1}, \ldots, f_{+c})$. U and V are both $k \times k$ matrices with orthonormal columns so that $U'U = V'V = I_k$, $D_\lambda = diag(\lambda_1, \ldots, \lambda_k)$, where λ_i is ith singular value.

For CA, we plot the first two columns of $A = D_r^{-1/2}UD_\lambda$ and $B = D_c^{-1/2}VD_\lambda$ (Greenacre and Hastie, 1987). The distances between two row (column) points are the approximated chi-square distance between profiles. Thus, we can interpret that the two closely located allele points tend to have similar frequencies. Eq. (1) can be re-expressed as

$$\chi^2 = \sum_i \sum_j \frac{(x_{ij} - x_{i+}x_{+j}/n)^2}{x_{i+}x_{+j}/n} = n\sum_i f_{i+}(r_i - c)'D_c^{-1}(r_i - c), \tag{3}$$

where $x_{i+} = \sum_j x_{ij}$, $x_{+j} = \sum_i x_{ij}$ and r_i is row profile such that $r_i = (x_{i1}/x_{i+}, \ldots, x_{ik}/x_{i+})'$. It means that χ^2/n is the weighted average of the squared chi-squared distances of the row profiles r_i to their centroid c. Thus, the significant χ^2 statistic can be geometrically interpreted as a significant deviation of the point from the origin. That is, if an allele is far from the origin, it implies that the allele contributes to break the Hardy–Weinberg equilibrium. Since the data matrix is symmetric, we do not need to obtain both A and B. Especially, $A = B$ if G is a positive semi-definite matrix.

For correspondence analysis biplot (CA biplot), let $A^* = UD_\lambda^\alpha$ and $B^* = VD_\lambda^{1-\alpha}$, where $D_\lambda^\alpha = diag(\lambda_1^\alpha, \lambda_2^\alpha, \ldots, \lambda_r^\alpha)$ and $0 \leqslant \alpha \leqslant 1$. It follows that $G = A^*B^{*\prime}$ which means $g_{ij} = a_i^{*\prime}b_j^*$, where g_{ij} is the (i, j)th element of G, $a_i^{*\prime}$ is the ith row of A and b_j^* is the first two columns of the jth row of B. It can be shown that (Gabriel, 1978)

$$G_{(s)} = \sum_{l=1}^{s} \lambda_l u_l v_l'$$

provides the best approximation of rank $s(\leqslant k)$ to G. That is, $g_{ij} = \sum_{l=1}^{k} \lambda_l u_{il} v_{jl} = a_i^{*\prime} b_j^*$ can be approximated by $g_{ij(s)} = \sum_{l=1}^{s} \lambda_l u_{il} v_{jl} = a_{i(s)}^{*\prime} b_{j(s)}^*$, $1 \leqslant s \leqslant k$, where u_{il}, v_{jl} are the (i, l)th and (j, l)th elements of U, V, respectively,

Download English Version:

https://daneshyari.com/en/article/418272

Download Persian Version:

https://daneshyari.com/article/418272

Daneshyari.com