# Construction of Two Locus Haplotype HLA-A,B Matrices For Poorly Studied Populations

Nina Fukshansky and K. Hummel

Prof. Dr. K. Hummel, Institut für Medizinische Mikrobiologie und Hygiene der Universität, Hermann-Herder-Str.11, D-7800 Freiburg

A genetic investigation of relationship should be based upon well defined frequency matrices for corresponding haplotypes. Two locus HLA-A,B matrices, however, exist for only a few thoroughly studied populations (Baur 1984). On the other hand some "similarities" between different populations have been well established (Nijenhuis 1984). Thus, attempts are made to use these similarities in order to perform analyses for populations whose matrices are only partial known. This can be done in different ways, depending on the definition of "similarity". We will show that different procedures within a large class result in the same approximate matrix. A regular method for constructing approximate matrices is proposed.

Let $\mathcal{M}$ be a well established two-locus haplotype matrix; the first locus A having alleles A(i) occurring with frequencies a(i)(i=0, 1,...,n), and the second locus B having alleles B(j) occurring with frequencies b(j)(j=0,1,...,m). The frequencies of haplotype A(i)B(j) will be designated as a(i,j). Then

$$\sum_{i=0}^{n} a(i,j) = b(j) \quad (j=0,1,\ldots,m); \qquad \sum_{j=0}^{m} a(i,j) = a(i) \quad (i=0,1,\ldots,n);$$

$$\sum_{i=0}^{n} a(i) = \sum_{j=0}^{m} b(j).$$

A corresponding matrix X for another insufficiently studied population is only partly known: all the allele frequencies $\bar{a}(i)$, $\bar{b}(j)$ (i=0,1,...,n; j=0,1,...,m) but only part of the haplotype frequencies x(i*,j*) are estimated. The problem is to compose the approximate matrix X with some fixed frequencies x(i*, j*) which obeys the limitations

$$\sum_{i=0}^{n} x(i,j) = \bar{b}(j) \quad j=0,1,\ldots,m; \qquad \sum_{j=0}^{m} x(i,j) = \bar{a}(i) \quad i=0,1,\ldots,n;$$

$$\sum_{j=0}^{m} \bar{b}(j) = \sum_{i=0}^{n} \bar{a}(i).$$

For this purpose we apply the so-called gravitational algorithm (known also as Deming-Stephan's or Shelekhovski's algorithm) (Pittel 1967) which has been used in problems of distribution of resources with limitations. At the first step the initial matrix $X_0 = \{x(i,j)^{(0)}\}$ of size n×m is chosen. The effect of variations

Advances in Forensic Haemogenetics 2
Edited by W.R. Mayr

on $x(i,j)^{(0)}$ will be discussed below. A sum of any line i in this matrix will not necessarily be equal to $\bar{a}(i)$. At the next step the matrix will be transformed to obtain this equality in each line: every element $x(i,j)^{(0)}$ will be multiplied

with $\bar{a}(i) / \sum_{j=0}^{m} x(i,j)^{(0)}$

to produce $x(i,j)^{(1)} = x(i,j)^{(0)} \bar{a}(i) / \sum_{j=0}^{m} x(i,j)^{(0)}$,

elements of matrix $X_1 = \left\{x(i,j)^{(1)}\right\}$. A sum of any column j in $X_1$ will not necessarily be equal to $\bar{b}(j)$. At the next step $X_1$ will be transformed to obtain this equality in each column: every element $x(i,j)^{(1)}$ will be multiplied with

$$\bar{b}(j) / \sum_{i=0}^{n} x(i,j)^{(1)}$$

to produce $x(i,j)^{(2)} = x(i,j)^{(1)} \bar{b}(j) / \sum_{i=0}^{n} x(i,j)^{(1)}$,

elements of matrix $X_2 = \left\{x(i,j)^{(2)}\right\}$. In $X_2$ the lines again should be improved etc. The subsequent improvements of lines and columns as described above establish an infinite sequence of matrices $X_k$ which, when $k \to \infty$, converges (Bregman 1967) to the limit matrix $X = \left\{x(i,j)\right\}$ maximizing the weighted entropy (Pittel 1967):

$$H = \sum_{i=0}^{n} \sum_{j=0}^{m} x(i,j) \ln \left[x(i,j)^{(0)} / x(i,j)\right]$$

under conditions

$$\sum_{i=0}^{n} x(i,j) = \bar{b}(j), \quad \sum_{j=0}^{m} x(i,j) = \bar{a}(i).$$

The matrix X (obviously) depends on the initial condition, i.e. on the choice of matrix $X_0$. Two different choices of $X_0$ seem to be reasonable:
1. As initial matrix $X_0$ the matrix $\mathcal{A}$ is used, $x^{(0)}(i,j) = a(i,j)$.
2. As initial matrix the disequilibrium matrix $Q = \left\{q(i,j)\right\}$ of matrix $\mathcal{A}$ with elements $q(i,j) = a(i,j) / \left[a(i) \cdot b(j)\right]$ is used.
Elements q(i,j) are equal to unity when alleles A(i) and B(j) are stochastically independent; otherwise q(i,j) are measures of dependence between A(i) and B(j). It has been proposed (Nijenhuis 1984) that in related populations disequilibria $\left\{q(i,j)\right\}$ will tend to a high degree of similarity.
We now show that both above mentioned choices of the matrix $X_0$ are equivalent, i.e. they both lead to the same limit matrix $X = \left\{x(i,j)\right\}$. Indeed, the solution maximizing

$$\tilde{H} = \sum_{i=0}^{n} \sum_{j=0}^{m} x(i,j) \ln \left[\tilde{x}(i,j)^{(0)} / x(i,j)\right]$$

where $\tilde{x}(i,j)^{(0)} = x(i,j)^{(0)} \cdot \alpha(i) \cdot \beta(j)$, maximizes also

$$H = \sum_{i=0}^{n} \sum_{j=0}^{m} x(i,j) \ln \left[ x(i,j)^{(0)} / x(i,j) \right] \quad \text{since}$$

$$H = \sum_{i=0}^{n} \sum_{j=0}^{m} x(i,j) \ln \left[ x(i,j)^{(0)} \alpha(i) \beta(j) / x(i,j) \right]$$

$$= \sum_{i=0}^{n} \sum_{j=0}^{m} x(i,j) \left\{ \ln \left| x(i,j)^{(0)} / x(i,j) \right| + \ln \alpha(i) + \ln \beta(j) \right\}$$

$$= \sum_{i=0}^{n} \sum_{j=0}^{m} x(i,j) \ln \left[ x(i,j)^{(0)} / x(i,j) \right]$$

$$+ \sum_{i=0}^{n} \ln \alpha(i) \sum_{j=0}^{m} x(i,j) + \sum_{j=0}^{m} \ln \beta(j) \sum_{i=0}^{n} x(i,j)$$

$$= H + \sum_{i=0}^{n} a(i) \ln \alpha(i) + \sum_{j=0}^{m} b(j) \ln \beta(j) = H + \text{const.}$$

In conclusion it should be noted that

1. if some elements $x(i^*,j^*)$ of the matrix X are fixed from the beginning, the corresponding elements $x(i^*,j^*)^{(0)}$ in $X_0$ should be nullified and $x(i^*,j^*)$ will be substracted from $\bar{a}(i^*)$ and $\bar{b}(j^*)$. After the application of the algorithm, the nullified elements with indices $(i^*,j^*)$ should be substituted for $x(i^*, j^*)$.
2. In HlA matrix alleles A(O), B(O) are collective terms for unprovable features. If not all of the $\bar{a}(i)$ and $\bar{b}(j)$ have been defined, then it is necessary to reduce the initial matrix $X_0$ to the corresponding size: if some $\bar{a}(i^*)$ ($\bar{b}(j^*)$) are not defined, the line $i^*$ (row $j^*$) of the matrix $X_0$ will be nullified and its elements will be substracted from the elements of the line O (row O).

References

Baur MP, Neugebauer M, Albert ED (1984) Reference tables of two-locus haplotype frequencies for all MHC marker loci. In: Albert ED, Baur MP, Mayr WR (eds) Histocompatibility Testing 1984. Springer, Berlin Heidelberg New York Tokyo, p 677-755

Bregman LM (1967) Proof of the convergence of G.V.Seleihov skii method for a problem with transportation constraints. Comput. Math and Math Phys 7: 191-204

Nijenhuis LE (1984) Estimation of two locus haplotype frequencies (in regional or local populations) from small population samples. Ann. Report 1984 Dr. Karl Landsteiner found. 33

Pittel BG (1967) A simple probability model of collective behavior. Probl Inform Transm (Problemy Peredachi Infor matsii) 3: 37-52

P.S. We produced HLA-A,B haplotype frequency matrices for more than 72 europid sub-populations and for 164 non-europid populations.