# Estimation of HLA haplotype frequencies in national or regional populations.

L.E.Nijenhuis, Central Laboratory of the Netherlands Red Cross Blood Transfusion Service, Amsterdam, the Netherlands

With about 20 HLA-A alleles, 40 B alleles and 10 C alleles, the total number of different HLA-A,B,C haplotypes amounts to 20 x 40 x 10= 8000. This implies that the greater part of the three-locus haplotypes will have low or even very low frequencies and that, even in rather large population samples, the estimates of these frequencies will have relatively considerable standard errors.
Even if those estimates could be obtained from direct countings of haplotypes, the sampling errors of most of them would be considerable. This is the more so where haplotype frequencies must be calculated from observed phenotypes, due to the multiple genotypical interpretations that are possible for each phenotype.
Each two-locus HLA-phenotype has at least two different genotypical possibilities; with three-locus phenotypes that number amounts to minimally four. These minimal numbers of genotypical possibilities are valid for full-house phenotypes, i.e. if visible heterozygosity exists for each of the two or three (A and B, or B and C, or A and B and C) genes involved in it. However, with part of the phenotypes, one or more of the genes may be represented by only one visible allele. Since one can never be sure whether such a situation should be explained as homozygosity for that allele, or rather as heterozygosity with a 'blanc' allele, the number of genotypical interpretations of certain phenotypes may be increased from 2 up to 5 for two-locus phenotypes, or from 4 up to 14 for three-locus phenotypes. And this situation greatly increases the standard errors of the estimated haplotype frequencies.
In comparison to the errors of the haplotype frequency estimates, those of the calculated gene frequencies of the separate HLA-A,B or C genes are relatively small.
Consequently, also the reliability of estimates of the haplotype frequencies could be assumed to be markedly improved if it were possible to calculate haplotype frequencies directly from gene frequencies.
It is, however, generally believed that such procedures are impossible, due to the wellknown linkage disequilibria between the separate HLA genes. Two preliminary studies have shown that that impossibility may appear less absolute than is generally believed.
The first study concerns two-locus haplotypes.
The results of the study suggest that, although gene-frequencies may vary between related populations, linkage disequilibrium relations are much more stable.
These equilibrium relations can be expressed as the quotient (Q) of the 'observed' haplotype frequency and its expectation (i.e. the product of the frequencies of the composing alleles):

$$Q_{i,j} = \frac{p(Ai,Bj)}{p(Ai).p(Bj)} \quad \text{or} \quad Q_{j,k} = \frac{p(Bj,Ck)}{p(Bj).p(Ck)}$$

Concerning a group of related populations (e.g. Caucasians), if a number of two-locus haplotypefrequency tables are available from the literature, mean Q values can be deduced from them:



Advances in Forensic Haemogenetics 1
Edited by B. Brinkmann and K. Henningsen

$$Q_{i,j}^{m} = \frac{N_1 \times Q_{i,j}^{1} + N_2 \times Q_{i,j}^{2} + \ldots\ldots + N_n \times Q_{i,j}^{n}}{N_1 + N_2 + \ldots\ldots + N_n}$$

in which $Q_{i,j}^{1}$, $Q_{i,j}^{2}$, etc. represent the Q values that are deduced from various haplotype frequency tables for a given HLA-A,B haplotype, and $N_1$, $N_2$ etc. are the sample sizes belonging to the various tables.

These mean Q values are used, together with the estimated gene frequencies of a sample from a local or regional population of limited size to compute the two-locus haplotype frequencies of that population:

$$p(Ai,Bj) = p(Ai).p(Bj).Q_{i,j}^{m} \quad \text{or} \quad p(Bj,Ck) = p(Bj).p(Ck).Q_{j,k}^{m}$$

This procedure may be assumed to yield relatively reliable two-locus haplotype frequency estimates, especially if the total size of the samples on which the various applied haplotype frequency tables were based, was considerably large.

Our second preliminary study concerned three-locus haplotype frequencies. It resulted into the conclusion that in most cases (for most of the B-alleles) three-locus haplotype frequency estimates can be calculated reasonably reliably by the formula

$$p(Ai,Bj,Ck) = \frac{p(Ai,Bj).p(Bj,Ck)}{p(Bj)}$$

For a few B-alleles the formula appeared unreliable; the most relevant of them concerns three-locus haplotypes with the B44 allele. Most probably this is due to the fact that B44 represents a so called 'broad specificity' that can be expected in the next future to be splitted up into a number of 'short specificities', each of them with its own linkage relations with certain HLA-A and C alleles.

Combination of the two aspects described above leads to the formula:

$$p(Ai,Bj,Ck) = p(Ai).p(Bj).p(Ck).Q_{i,j}^{m}.Q_{j,k}^{m}$$

Especially with regional population samples of limited size, and with two-or three-locus haplotypes of rather low frequency, the results with the calculation methods that are described above may be assumed to be much more reliable than the estimates that would be deduced from the distribution of the phenotypes in the sample.