Arch. Tierz., Dummerstorf 44 (2001) 2, 231-236

[1]Research Institute of Animal Production, Prague-Uhříněves, Czech Republic;
[2]Research Institute for Biology of Farm Animals, Dummerstorf, Germany
[3]Research Institute of Animal Production, Nitra, Slovak Republic

JOCHEN WOLF[1], GERHARD DIETL[2], DANA PEŠKOVIČOVÁ[3], DIETER SUMPF[2]
and MARTINA LANGHAMMER[2]

Heterozygosity between populations - a possible alternative to measures of genetic distance

Dedicated to Professor Dr. Erhard Kallweit on the occasion of his 65th birthday

Summary

It is shown that the heterozygosity between two populations defined as the expected proportion of heterozygotes in their cross can be used as a complement to measures of genetic distance. This new measure has favourable mathematical properties (fulfils the triangular inequality) and can be well interpreted from the biological point of view. Its main importance will be in comparing (potential) crosses amongst each other.

Key Words: genetic distance, heterozygosity, distance measures

Zusammenfassung

Titel der Arbeit: **Heterozygotiegrad zwischen Populationen - eine mögliche Alternative zu genetischen Abstandsmaßen**

Es wird gezeigt, dass der Heterozygotiegrad zwischen zwei Populationen, der definiert ist als der erwartete Anteil von Heterozygoten in ihrer Kreuzung, eine mögliche Alternative für genetische Abstandsmaße ist. Dieses neue Maß hat günstige mathematische Eigenschaften (genügt der Dreiecksungleichung) und kann unter biologischem Gesichtspunkt gut interpretiert werden. Seine Bedeutung dürfte hauptsächlich im gegenseitigen Vergleich (potentieller) Kreuzungskombinationen liegen.

Schlüsselwörter: Genetischer Abstand, Heterozygotiegrad, Abstandsmaße

Introduction

Several measures were proposed for calculating the genetic distance between two populations (GREGORIUS 1974, 1984; NEI, 1972; PREVOSTI et al., 1975; ROGERS, 1972). These measures were derived from different approaches. NEI's standard genetic distance (NEI, 1972) is based on a hypothesis of the evolution process. He assumes a linear relationship between the genetic distance and the time in evolution the two genotypes were separated from each other. GREGORIUS (1974, 1984) stated a set of conditions to which a distance measure must comply and derived his distance measure from this basis. The same distance measure was used by PREVOSTI et al. (1975). From the animal breeders' point of view, a measure between two populations related to the expected proportion of heterozygotic individuals in the resulting cross, is appealing. In the following text it will be shown that the heterozygosity between populations is such a measure.

Definition of the Heterozygosity Between Populations

Let p_{Xij} and p_{Yij} be the frequency of the jth allele at the ith locus in populations X and Y, respectively. Then, when crossing populations X and Y by mating at random, the expected frequency of homozygotes at locus i, g_{XYi}, is:

$$g_{XYi} = \sum_{j=1}^{n_i} p_{Xij} p_{Yij}$$

where n_i is the number of alleles at locus i.

The expected frequency of heterozygotes at locus i in the cross, h_{XYi}, is then:

$$h_{XYi} = 1 - g_{XYi} = 1 - \sum_{j=1}^{n_i} p_{Xij} p_{Yij} \qquad [1]$$

This quantity is formally similar to the heterozygosity in a population (for X=Y). We call it "heterozygosity between populations at locus i" therefore. This measure has similar properties as a distance measure. It can be well interpreted from the biological point of view and has favourable properties from the mathematical point of view. The heterozygosity between populations at locus i can be estimated by replacing the allele frequencies in [1] by its estimates.

The variance of the heterozygosity between populations is (NEI and ROYCHOUDHURY, 1974):

$$var(\hat{h}_{XYi}) = \frac{1}{m_X m_Y}\left\{(1-m_X-m_Y)\left(\sum_{j=1}^{n_i}\hat{p}_{Xij}\hat{p}_{Yij}\right)^2 + (m_X-1)\sum_{j=1}^{n_i}\hat{p}_{Xij}^2\hat{p}_{Yij} + (m_Y-1)\sum_{j=1}^{n_i}\hat{p}_{Xij}\hat{p}_{Yij}^2 + \sum_{j=1}^{n_i}\hat{p}_{Xij}\hat{p}_{Yij}\right\}$$

with m_X and m_Y being the number of individuals from populations X and Y, respectively.

For a given set of loci, an average heterozygosity between populations (H_{XY}) can be defined. Its estimate can be calculated simply as arithmetic mean of the estimates of the heterozygosities at the individual loci:

$$\hat{H}_{XY} = \left(\sum_{i=1}^{r}\hat{h}_{XYi}\right) / r$$

with the variance

$$var(\hat{H}_{XY}) = \sum_{i=1}^{r}(\hat{h}_{XYi} - \hat{H}_{XY})^2 / [r(r-1)]$$

where r is number of loci in this set.

Properties of the Heterozygosity Between Populations

First consider the special case of two alleles at locus i. Putting for simplicity $p = p_{Xi1}$ and $q = p_{Yi1}$, the equation for the heterozygosity between population X and Y reduces to

$$h_{XYi} = p + q - 2pq$$

For p = constant, h_{XYi} is a linear function of q only. For $p = 0.5$, h_{XYi} is independent of q

and takes the value 0.5. h_{XYi} takes its minimal value of zero for $p = q = 0$ and $p = q = 1$. The maximal value of 1 is reached for the combination $p = 0$ and $q = 1$ or vice versa ($p = 1$ and $q = 0$).

In the general case of n_i alleles at the ith locus, n_i being any positive integer, the lower bound of h_{XYi} is zero and its upper bound is unity. The lower bound is reached if and only if in both populations the same allele occurs with the frequency one (i.e. all other alleles have the frequency zero in both populations). The upper bound is reached if and only if both populations do not have any allele in common.

The measure h_{XYi} fulfils the triangular inequality:

$$h_{XZi} \leq h_{XYi} + h_{YZi} ,$$

where X, Y and Z are three populations and the h's are the heterozygosities referring to the appropriate pairs of populations. The proof is given in the Appendix.

If all allele frequencies at locus i in the population X are equal, i.e. $p_{Xij} = 1/n_i$ for all j, then equation [1] simplifies to

$$h_{XYi} = 1 - \frac{1}{n_i} \sum_{j=1}^{n_i} p_{Yij} = 1 - \frac{1}{n_i}$$

as the sum is unity. In that case h_{XYi} is independent of the allele frequencies in the population Y.

Another interesting case is when in population X only one allele, say j^*, is present, i.e. $p_{Xij^*} = 1$. Then equation [1] reduces to

$$h_{XYi} = 1 - p_{Yij^*}$$

That means, h_{XYi} depends only on the frequency of the same allele in population Y and is independent on the remaining allele frequencies in Y. For completeness it should be added that $h_{XYi} = h_{YXi}$, so that X and Y can be exchanged without changing the result.

Discussion

GREGORIUS (1974) stated four conditions for a measure to be a distance measure: (i) it takes only nonnegative values, (ii) it is symmetric (the distance between A and B is the same as the distance between B and A), (iii) it takes the value zero if and only if both populations are identical, (iv) it fulfils the triangular inequality. The heterozygosity between populations meets conditions (i), (ii) and (iv), but not (iii). Therefore this measure has some similarity to a distance measure, but is no distance measure in the sense of the above definition. The explanation for the divergence from a distance measure with respect to condition (iii) is given below.

To illustrate the differences between several measures of genetic distance and the heterozygosity between populations, the Table gives numeric values for some basic situations. Only one locus is considered. All animals in the hypothetic populations X and Y are assumed to be identical or both populations are assumed to consist of one animal each. Capital letters A, B, C and D designate different alleles. The numeric values of the new measure given by equation [1] are most similar to the values of GREGORIUS' distance (GREGORIUS, 1974, 1984). Similarly as in NEI's standard genetic distance (NEI, 1972), the situations AA AB and AB BC are discriminated against by the new measure.

As already stated above (divergence from condition (iii) of a distance measure), the heterozygosity between populations may differ from zero, though the genotype and the allele frequencies in both populations will be equal (situation AB AB in the Table). At first glance, it seems to be illogical that a value greater than zero is calculated although the two populations X and Y are identical from a genetic point of view. But when comparing X and Y on the basis of alleles and not genotypes, half of the comparisons will yield equal alleles (A_x - A_Y, B_x - B_Y, read A_x as "allele A from population X" etc.) and half of the comparisons unequal alleles (A_x - B_Y, A_Y - B_x). This is the way the new measure is defined and this is the basic difference from the distance measures. The heterozygosity between populations is defined in respect to what will happen when the populations are crossed amongst each other and contains implicitly a dynamic aspect.

Table
Comparison between several measures of genetic distance and the heterozygosity between populations

Genotypes X Y	GREGORIUS 1984	NEI 1972 std.+max.	NEI 1972 min.	ROGERS 1972	Hetero-zygosity
AA AA	0	0	0	0	0
AA AB	0.50	0.35	0.25	0.50	0.50
AA BB	1	∞	1	1	1
AB AB	0	0	0	0	0.50
AA BC	1	∞	0.75	0.87	1
AB BC	0.50	0.69	0.25	0.50	0.75
AB CD	1	∞	0.50	0.71	1

NEI's minimum genetic distance (NEI, 1972) is defined as

$$d_{min} = \frac{1}{2}\sum_{j=1}^{n_l}(p_{Xlj} - p_{Ylj})^2 = \frac{1}{2}\left(\sum_{j=1}^{n_l}p_{Xlj}^2 + \sum_{j=1}^{n_l}p_{Ylj}^2\right) - \sum_{j=1}^{n_l}p_{Xlj}p_{Ylj}$$

When assuming Hardy-Weinberg equilibrium in the populations X and Y this definition can be rewritten to

$$d_{min} = h_{XYl} - \frac{1}{2}(h_{Xl} + h_{Yl})$$

where h_{Xl} and h_{Yl} are the heterozygosities within the populations X and Y, respectively, calculated under the above assumption. NEI's minimum genetic distance can therefore be interpreted as an increase in heterozygosity when crossing populations X and Y which are in Hardy-Weinberg equilibrium. This measure may be useful when comparing crosses with purebred populations, but for comparing crosses among each other the uncorrected heterozygosity between populations as defined in equation [1] should be preferred. ROGERS' distance (ROGERS, 1972) is the square root from NEI's minimum genetic distance and therefore related to the distance measure [1] in a similar way.

As in studies with microsatellites all alleles from the given set of loci under consideration can be identified in general, the estimates of heterozygosities within the populations can be calculated by counting the number of heterozygotes and relating it to the overall number of animals. Therefore NEI's minimal genetic distance could be modified in such a way that h_{Xl} and h_{Yl} are replaced by the heterozygosities calculated in the more direct way by counting heterozygotes and not from allele frequencies assuming Hardy-Weinberg equilibrium. This might yield more precise estimates of the increase of the proportion of heterozygotes in potential crosses, but on the other hand, could yield negative estimates of genetic distances, when the allele distribution in populations X and/or Y is far from Hardy-Weinberg equilibrium.

The heterozygosity between populations should be mainly used when the degree of heterozygosity of potential crosses is of interest. For other purposes such as clustering of genotypes with no respect to their future use in crossbreeding programs the absolute genetic distance of GREGORIUS (1984) may be more suitable and for investigations related to the evolutionary process NEI's standard genetic distance (NEI, 1972) will be the method of choice. NEI's minimal genetic distance (NEI, 1972) and ROGERS' distance (ROGERS, 1972) have the unfavourable property that, in certain situations, they do not give maximal values, even though both populations do not share common alleles (combination AB CD in Table 1). They should therefore be used with care.

Acknowledgements

The research project was supported by the grant EP7123 of the Grant Agency of the Ministry of Agriculture of the Czech Republic and by the research project MZE-M02-99-02 of the Research Institute of Animal Production Prague-Uhríneves. J. Wolf thanks the Bundesminsterium für Ernährung, Landwirtschaft und Forsten der Bundesrepublik Deutschland for the financial support of several stays at the Research Institute for the Biology of Farm Animals, Dummerstorf.

References

GREGORIUS, H.R.:
Genetischer Abstand zwischen Populationen. I. Zur Konzeption der genetischen Abstandsmessung. Silv. Genet. 23 (1974), 22-27

GREGORIUS, H.R.:
A unique genetic distance. Biom. J. 26 (1984), 13-18

NEI, M.:
Genetic distance between populations. Am. Naturalist 106 (1972), 283-292

NEI, M.; ROYCHOUDHURY, A.K.:
Sampling variances of heterozygosity and genetic distance. Genetics 76 (1974), 379-390

PREVOSTI, A.; OCAÑA, J.; ALONSO, G.:
Distances between populations of *Drosophila subobscura*, based in chromosome arrangement frequencies. Theor. Appl. Genet. 45 (1975), 231-241

ROGERS, I.S.:
Measures of genetic similarity and genetic distance. Stud. Genet. 7 (1972), 145-153

WEIR, B.S.:
Genetic Data Analysis. Methods for Discrete Population Genetic Data, Sinauer Associates, Inc., Publishers, Sunderland, Massachusetts, 1990

Appendix: Proof of the Triangular Inequality

Consider any locus i and let n be the number of alleles. For simplicity, p_j, q_j and r_j is written instead of p_{Xij}, p_{Yij} and p_{Zij} for the frequency of the jth allele ($j = 1, 2, ..., n$) in populations X, Y and Z, respectively. Similarly, i is omitted in the index of the heterozygosities at locus i.

The heterozygosities can alternatively be written in the following form:

$$h_{XY} = \sum_{j=1}^{n} q_j(1-p_j),\ h_{YZ} = \sum_{j=1}^{n} r_j(1-q_j),\ h_{YZ} = \sum_{j=1}^{n} r_j(1-p_j)\ . \qquad [A1]$$

It is to be shown that

$$h_{XZ} \leq h_{XY} + h_{YZ}$$

or

$$\Delta h = h_{XY} + h_{YZ} - h_{XZ} \geq 0 \, . \qquad [A2]$$

Inserting [A1] to [A2] it is obtained:

$$\Delta h = \sum_{j=1}^{n}\left[q_j(1-p_j) + r_j(1-q_j) - r_j(1-p_j)\right] = \sum_{j=1}^{n}\Delta h_j$$

with

$$\begin{aligned}\Delta h_j &= q_j(1-p_j) + r_j(1-q_j) - r_j(1-p_j)\\ &= q_j(1-p_j) + r_j(p_j-q_j)\end{aligned}$$

where $0 \leq p_i, q_i, r_i \leq 1$.

For showing that $\Delta h \geq 0$, it is sufficient to show that $\Delta h_j \geq 0$ for any j. Consider first the case that $p_i \geq q_j$, r_j taking any value. Then

$$\Delta h_j = q_j(1-p_j) + r_j(p_j-q_j) \geq 0 \, .$$

Assume now that $p_j < q_j$ and $r_j \geq q_j$. Then

$$\begin{aligned}\Delta h_j &= q_j(1-p_j) + r_j(p_j-q_j)\\ &= q_j(1-r_j) + p_j(r_j-q_j) \geq 0 \, .\end{aligned}$$

The last case to consider is $p_j < q_j$ and $r_j < q_j$. It is:

$$\begin{aligned}\Delta h_j &= q_j(1-p_j) + r_j(1-q_j) - r_j(1-p_j)\\ &= (q_j-r_j)(1-p_j) + r_j(1-q_j) \geq 0 \, .\end{aligned}$$

Herewith it has been shown that the triangular inequality is valid for any allele frequencies. Because of the additivity of the heterozygosity between populations in respect to the loci the triangular inequality holds not only for the heterozygosity at a given locus, but for the average heterozygosity as well.

Received: 2000-06-21

Accepted: 2000-10-26

Authors' addresses
Dr. JOCHEN WOLF
Research Institute of Animal Production (VÚŽV)
P.O.BOX 1
CZ 10401 Praha 114 - Uhříněves
Czech Republic

Dr. GERHARD DIETL, Dr. DIETER SUMPF, Dr. MARTINA LANGHAMMER
Research Institute for Biology of Farm Animals
Wilhelm-Stahl-Allee 2
D - 18196 Dummerstorf
Germany

DANA PEŠKOVIČOVÁ
Research Institute of Animal Production
Hlohovská 2
SK 94992 Nitra
Slovak Republic