Acta Biotheor (2011) 59:213–229
DOI 10.1007/s10441-011-9138-9

REGULAR ARTICLE

# The Analysis of Association Between Traits When Differences Between Trait States Matter

**Hans-Rolf Gregorius**

Received: 19 July 2010 / Accepted: 18 June 2011 / Published online: 28 June 2011


**Abstract** Because of their elementary significance in almost all fields of science, measures of association between two variables or traits are abundant and multiform. One aspect of association that is of considerable interest, especially in population genetics and ecology, seems to be widely ignored. This aspect concerns association between complex traits that show variable and arbitrarily defined state differences. Among such traits are genetic characters controlled by many and potentially polyploid loci, species characteristics, and environmental variables, all of which may be mutually and asymmetrically associated. A concept of directed association of one trait with another is developed here that relies solely on difference measures between the states of a trait. Associations are considered at three levels: between individual states of two variables, between an individual state of one variable and the totality of the other variable, and between two variables. Relations to known concepts of association are identified. In particular, measures at the latter two levels turn out to be interpretable as measures of differentiation. Examples are given for areas of application (search for functional relationships, distribution of variation over populations, genomic associations, spatiogenetic structure).

**Keywords** Measures of association · Asymmetry of association · Levels of association · Variable state differences · Complex traits · Dissociation · Shift transformation · Measure of differentiation · Functional relationship · Genomic associations · Ecological genetics

H.-R. Gregorius (✉)
Abteilung Forstgenetik und Forstpflanzenzüchtung, Universität Göttingen, Büsgenweg 2, 37077 Göttingen, Germany
e-mail: hgregor@gwdg.de

H.-R. Gregorius
Institut für Populations- und ökologische Genetik, Am Pfingstanger 58, 37075 Göttingen, Germany

## 1 Introduction

The detection of kinds and degrees of relationship, or association, between two traits is one of the most fundamental issues in scientific research. Accordingly numerous are the measures that have been proposed. Typically, associations are determined between two variables (or traits) of the same type, in the sense that both variables are discrete (qualitative, categorical, etc.), both ordinal, or both continuous, for example. Associations between variables of different type are usually treated by transforming one type into the other. With one variable discrete and the other continuous, for example, the latter is usually partitioned more or less arbitrarily into classes, making both variables discrete. While for discrete variables association is usually determined by measures of deviation from stochastic independence, association between real-valued or ordered (ordinal) variables is commonly treated in terms of measures of covariation and thus of monotonicity of relationships between the variables (the classical papers of Goodman and Kruskal compiled in Goodman and Kruskal 1979, and the monography of Liebetrau 1983, still provide a suitable overview of the most common measures). More recent approaches are based on measuring the dependence of the distribution of one variable on the distribution of a second variable with the help of the variance of conditional probabilities (see e.g. Hsing et al. 2005; or Liu 2005). In these approaches variables are allowed to be of different type. The resulting indices again measure the deviation from stochastic independence but in an asymmetrical way. They are measures of directed association and are thus applicable to analyses of cause-effect relationships.

Complex variables in particular are frequently characterized by variable differences between their states, where the applied measures of difference may be of quite diverse kind. Consideration of variable differences in analyses of association introduces a perception that cannot be captured simply by methods of covariation or of transformation of joint frequency distributions. Despite its obviousness, this perception seems to have attracted little, if any, explicit attention in association studies. As an example, in a biological context, variable differences are essential whenever problems of differential relatedness or similarity of species, populations or individuals are addressed in connection with the environmental conditions in which they are found or to which they are presumably adapted. The systematic, genetic or phylogenetic traits of these entities are mostly multidimensional, as are most environmental characteristics of interest. The currently popular genome-scale studies in phylogenetics and ecological genetics pose particularly obvious challenges in this respect. With the exception of rare events of perfect cloning, each genotype identified at a genomic level is unique (realized by only one individual). The plain fact that genotypes are not identically repeated thus precludes any classical analysis of association of the genetic trait with other functional, phenotypic, or ecological traits (see e.g. Hughes 2008, for a recent commentary).

Nevertheless, genotypes are composed of gene-types (alleles) that may be shared among individuals. Genotypes thus differ to variable degrees, and these differences may be associated with certain differences that are measurable between the states of other traits or variables. Actually, the detection of variation *per se* relies on the

ability to discriminate and thus to recognize differences. For qualitative traits, for example, differences are measured in a binary fashion by stating sameness and its opposite (a value of 0 indicates sameness of two individuals and 1 indicates their differentness). Variation thus becomes visible through differences, which suggests that studies of association between variables should first and foremost take differences between the states of a variable into consideration. Herewith, neither the type of variable (e.g. qualitative or quantitative), the combination of types of variables nor the way differences between the states of the variables are measured should impose any restrictions on the analyses. The present paper is devoted to the elaboration of a conceptual approach to the assessment of association that takes all of these aspects into consideration.

## 2 Preliminary Deliberations

When two independently specified features are observed, their association is basically determined by those population members that display both features. The features commonly appear as states of two traits (or variables) $X$ and $Y$, say. The more members of state $x$ of $X$ that also hold state $y$ of trait $Y$, and the more distinctly the members not holding state $y$ differ from $x$, the more strictly can state $y$ be considered to be associated with state $x$. As becomes evident from this formulation, considerations of association are of an intrinsically *directed* nature, and this reflects an essential prerequisite for the detection of cause-effect relationships in the sense that $y$ (effect) is determined by $x$ (cause). If not stated otherwise, the term "population" will be conceived in a wide sense as any specified finite or infinite collection of objects.

Perfect (or strict) association is thus characterized by two conditions, one of which requires that all $x$-members display $y$, and the other requires that the $X$-states of all members that do not hold state $y$ be distinct from $x$. The second condition reflects the expectation that even if the possession of $x$ would always imply possession of $y$, the association would not be considered perfect if the $X$-states of members not holding $y$ could come close to state $x$. In that case the association would become imperfect because of insufficient separation of state $x$ from members not holding state $y$. In fact, the second condition implies the first, since if all members not holding $y$ differ from $x$, then all $x$-members, if there are any, show state $y$. At the other extreme, $y$ can be viewed to be perfectly dissociated from $x$ if all $y$-members distinctly differ from $x$ in their $X$-states. Loosely speaking, association of $y$ with $x$ is strong if members that do not hold state $y$ differ distinctly in their $X$-states from $x$, and dissociation is strong if members that do hold state $y$ differ distinctly in their $X$-states from $x$.

Intermediate situations arise when among $y$-members the distribution of differences of $X$-states from $x$ is similar to that for members of the remainder of the population. Complete absence of association of $y$ with $x$ would then be stated if the possession of $y$ is not implied in any particular way by the possession of $x$ or by its differences from other $X$-states. The same situation represents the absence of dissociation of $y$ from $x$. This case is realized if the distribution of $X$-states among

*y*-members is the same as in the remainder of the population. In other words, trait *X* varies (stochastically) independently of state *y*. One concludes from these deliberations that the assessment of dissociation and association basically depends on the distribution of differences of *X*-states from *x* among *y*-members and within the remainder of the population, respectively. Variable differences between *X*-states thus imply that all *X*-states (i.e. not just state *x*) are potentially involved in the assessment of association of *y* with *x*.

To illustrate the transformations creating association, let $\Omega_y$ denote the set of *y*-members and let $\Omega_y^c$ denote its complement (the remainder). By the above explanations, the association of *y* with *x* increases as members from $\Omega_y^c$ become more distinct from state *x* for their *X*-trait. This is achieved by a *frequency shift* within $\Omega_y^c$ such that the frequency of an *X*-state that differs more from *x* than another *X*-state is increased, and in return the frequency of the other state is decreased by the same amount. Starting from a situation in which the distribution of *X* is the same in $\Omega_y$ and $\Omega_y^c$ (and association is thus absent), such shift transformations can be applied to create any difference between the two distributions and to observe the resulting gains or losses in association (see Fig. 1 for an illustration).

Therefore, in order to assess the associations realized in a particular joint distribution of *Y* and *X*, it is meaningful to consider the overall frequency shifts among *X*-states within $\Omega_y^c$ that are required to transform the distribution of *X* within $\Omega_y^c$ into the distribution of *X* within $\Omega_y$. The sum of frequency shifts quantifies the deviation from stochastic independence and thus, in concert with the pertaining differences from *x*, determines the degree of association. An assessment of dissociation can be achieved analogously by carrying out frequency shifts within $\Omega_y$ in order to match the distribution of *X* within $\Omega_y^c$. This approach will be detailed later on in the appropriate sections. Note that because of the directedness of association, difference measures are relevant only between the states of one trait unless the reverse association is additionally taken into account.

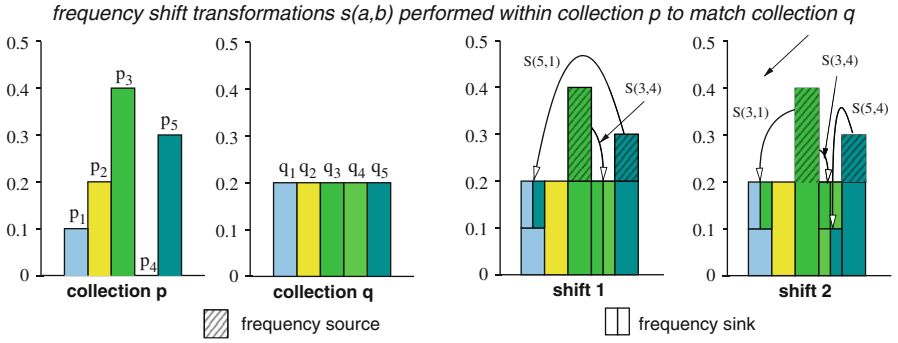


**Fig. 1** Two frequency shift transformations performed within collection **p** (corresponding to $\Omega_y^c$) in order to match the frequency distribution in collection **q** (corresponding to $\Omega_y$). Differences between states are supposed to correspond to differences in *color*. On this basis, shift transformation 1 represents a smaller total change in trait state (*color*) than shift transformation 2 (adapted from Gillet et al. 2004)

The above explanations show that difference measures are required for that trait with which the other trait is considered to be associated. Of course, all of these and of the following considerations apply identically when the determination of the two traits is reversed (association of $x$ with $y$). In this way, analyses of association can be conducted in both directions with the aim of detecting asymmetries that hint at cause-effect relations, for example. Since the present paper is specifically concerned with the effect of variable differences between trait states on the assessment of association, comparability of associations in both directions must rely on comparability of the difference measures applied to the involved traits.

## 3 Properties of Difference Measures

The characterization of situations of perfect association and perfect dissociation both depend on the notion of distinctness of trait states. In fact, perfect association or dissociation cannot be realized unless proper meaning is given to ideas of perfect or complete distinctness, differentness, or dissimilarity. If differences are bounded from above, their maximum value can be conceived of as specifying the situation of complete dissimilarity. Yet, even if the differences can become arbitrarily large, it may be meaningful to draw a threshold beyond which trait states are regarded to be completely dissimilar. Indeed, if differences may become infinitely large, statements as to the strength of dissociation become arbitrary. Hence, since the measurement of association or dissociation does not depend on scale and operates relative to the range of realizable differences, it is reasonable to consider only bounded difference measures that preferably vary between 0 and 1.

If from the outset a measure of difference is subject to a limitation with pertinent interpretation (such as "complete" differentness, distinctness, dissimilarity or separation), the appropriate normalization would be $\hat{d} = d/u$, where $u$ is the maximum value of $d$. Otherwise, if the maximum value $u$ is not predefined but has a specific meaning, the appropriate normalization might be $\hat{d} = d/u$ for $d \leq u$ and $\hat{d} = 1$ for $d \geq u$. This normalization may however be unsatisfactory if $d$ may reach very large values that are realized only in exceptional cases. In such cases the normalization

$$\hat{d} := \begin{cases} \frac{d}{1+d} \cdot \frac{e^{u-1}+d}{e^{u-1}+u} \cdot \frac{1+u}{u} & \text{if} \quad 0 \leq d \leq u \\ 1 & \text{if} \quad d \geq u \end{cases}$$

is more appropriate. The normalization is a strictly increasing function of $d$ that is convex for $u < 1$, is concave for $u > 1$, and attains its maximum value of 1 exactly for $d \geq u$. If $u = 1$ then $\hat{d} = d$ for $d \leq u$. As $u \to \infty$ one obtains $\hat{d} = \frac{d}{1+d}$. Note that all of the above normalizations decrease for fixed $d$ when the threshold value $u$ is increased ($u \geq d$). This rules out intersection of the functions $\hat{d}$ of $d$.

Given this connection between bounded measures of difference and perfect association as well as opportunities for appropriate normalization, the following developments will solely refer to difference measures that vary over the unit interval. The extremal values 0 and 1 are interpreted as complete similarity and

complete dissimilarity, respectively. This understanding conforms with the above requirement that difference measures for different traits have to be comparable in order to allow associations to be considered in both directions and by this to allow for the detection of asymmetries in association.

## 4 State-to-State Level of Association

It is argued above that the assessment of association or dissociation should be based on frequency shift transformations that are performed on the distribution of a variable in one set of objects with the aim of matching the distribution of that variable in another set of objects. The two sets are $\Omega_y$ (the set of objects with trait state $Y = y$) and its complement $\Omega_y^c$ (where $Y \neq y$), and the frequency shifts are performed on the distribution of trait $X$ in either of the two sets to match the distribution in the other set. With each shift, the frequencies of the differences between the involved $X$-states and a reference state $x$ are changed. This was shown to directly determine changes in association (or dissociation) of state $y$ of trait $Y$ with (from) state $x$ of trait $X$. In order to develop a measure of association from these principles, the notation listed in Table 1 is also needed.

Frequency shifts are carried out from $X$-states that are more frequent in $\Omega_y^c$ than in $\Omega_y$ to $X$-states that are less frequent in $\Omega_y^c$ than in $\Omega_y$. The former states will be called frequency sources and the latter frequency sinks. Matching the distribution of $X$ in $\Omega_y^c$ with the distribution in $\Omega_y$ thus requires that parts of the frequency sources be removed and added (shifted) to the sink frequencies. Herewith, state $X = a$ is a frequency source or frequency sink within $\Omega_y^c$ according to whether $P(X = a \mid Y \neq y) - P(X = a \mid Y = y)$ is positive or negative, respectively. Frequency shifts $\sigma(a, b)$ from frequency source states $X = a$ to frequency sink states $X = b$ within $\Omega_y^c$ are required in order to transform the distribution of $X$ within $\Omega_y^c$ into the distribution of $X$ within $\Omega_y$. A shift transformation must therefore level out the differences between frequency sources and sinks (see Fig. 1 for an illustration), i.e. $\sigma(a,b) \geq 0$ for all $a$ and $b$, and

**Table 1** Notation

| |
|---|
| $d_X(a, b)$:= measure of difference between the states $a$ and $b$ of trait $X$ |
| $P(\Gamma) :=$ relative frequency (in finite populations or samples) or probability of an event $\Gamma$; e.g. $P(X = a)$ is the relative frequency of state $a$ of trait $X$ |
| $P(\Gamma_1 \mid \Gamma_2) :=$ conditional probability of event $\Gamma_1$ given event $\Gamma_2$; e.g. $P(X = x \mid Y \neq y)$ is the frequency of population members holding state $x$ among members not holding state $y$ |
| $E(Z)$ := average or expectation of variable $Z$; e.g. $E(d_X(X, a)) = \sum_b d_X(a, b) \cdot P(X = b)$ is the average difference of trait $X$ from state $a$ (where $Z = d_X(X, a)$) |
| $E(Z \mid \Gamma) :=$ conditional expectation of variable $Z$ given event $\Gamma$; e.g. $E(d_X(X, a) \mid Y = y)$ is the average difference of trait $X$ from state $a$ among members holding state $y$ of trait $Y$ |

For continuous variables, the probabilities have to be replaced by probability densities, and the expectations appear as integrals.

$$\sum_b \sigma(a,b) = P(X=a \mid Y \neq y) - \min\{P(X=a \mid Y \neq y), P(X=a \mid Y=y)\},$$
$$\sum_a \sigma(a,b) = P(X=b \mid Y=y) - \min\{P(X=b \mid Y \neq y), P(X=b \mid Y=y)\}.$$

Any frequency shift may increase or decrease the association depending on the size of the difference of the source state and of the sink state from $x$. The gain or loss in association is therefore determined by the size of the frequency shift and by the difference between the two state differences. In particular, if both differences are equal, association is unaffected by the shift; if the difference of $x$ from the source state exceeds that from the sink state, association is increased, and it is decreased otherwise. Hence, summing up all individual shifts leads to an overall net gain or loss of association according to

$$T_{y,x}(\sigma) := \sum_{a,b} \sigma(a,b) \cdot [d_X(a,x) - d_X(b,x)] \tag{1}$$

There may be many shift transformations that fulfill the above conditions. However, as follows directly from the shift characteristics,

$$T_{y,x}(\sigma) = \sum_a [P(X=a \mid Y \neq y) - P(X=a \mid Y=y)] \cdot d_X(a,x)$$

which shows that $T_{y,x}(\sigma)$ does not depend on which shift transformation $\sigma$ is applied.

The sum in the representation of $T_{y,x}$ can be decomposed into two sums, the first of which equals the average difference $E(d_X(X,x) \mid Y \neq y)$ of $X$-states from $x$ within $\Omega_y^c$, and the second equals the corresponding average $E(d_X(X,x) \mid Y=y)$ within $\Omega_y$. One thus obtains the following two equivalent representations of $T_{y,x}$:

$$\begin{aligned} T_{y,x} &= E(d_X(X,x) \mid Y \neq y) - E(d_X(X,x) \mid Y=y) \\ &= \frac{E(d_X(X,x) \mid Y \neq y) - E(d_X(X,x))}{P(Y=y)} \end{aligned}$$

where $E(d_X(X, x))$ is the overall average difference of $X$-states from $x$ [recall that $E(d_X(X,x)) = E(d_X(X,x) \mid Y \neq y) \cdot P(Y \neq y) + E(d_X(X,x) \mid Y=y) \cdot P(Y=y)$]. With the help of this expression, the absence of association, its presence, and the presence of dissociation can be stated as $T_{y,x} = 0$, $T_{y,x} > 0$, and $T_{y,x} < 0$, respectively.

Moreover, $E(d_X(X,x) \mid Y \neq y) = \sum_a P(X=a \mid Y \neq y) \cdot d_X(a,x) \leq 1$ with equality only if for each $a$ with $P(X=a \mid Y \neq y) > 0$ one has $d_X(a,x) = 1$, and $P(X=x \mid Y \neq y) = 0$. This condition conforms precisely with the definition of perfect association. It suggests normalization of $T_{y,x}$ so as to yield a measure of association of $y$ with $x$ with upper limit equal to 1. There are two ways to normalize depending on whether the first or the second representation of $T_{y,x}$ is used. For reasons of comparison with existing measures, the second representation will be given preference. This normalization yields the *measure of association of y with x*

$$\mathcal{A}^{+}(Y=y\,|\,X=x)=\frac{E(d_X(X,x)\,|\,Y\neq y)-E(d_X(X,x))}{1-E(d_X(X,x))}$$

By symmetry of arguments (and since perfect dissociation is realized only for $E(d_X(X,x)\,|\,Y=y)=1$), one obtains the pertaining *measure of dissociation of y from x* as

$$\mathcal{A}^{-}(Y=y\,|\,X=x)=\frac{E(d_X(X,x)\,|\,Y=y)-E(d_X(X,x))}{1-E(d_X(X,x))}$$

The two measures are connected by

$$\mathcal{A}^{+}(Y=y\,|\,X=x)\cdot P(Y\neq y)+\mathcal{A}^{-}(Y=y\,|\,X=x)\cdot P(Y=y)=0$$

Both measures can be combined into a single measure that varies from $-1$ to $+1$, indicating association by positive values, its absence by zero, and dissociation by negative values:

$$\begin{aligned}\mathcal{A}(Y=y\,|\,X=x)&=\iota\cdot\max\{\mathcal{A}^{+}(Y=y\,|\,X=x),\,\mathcal{A}^{-}(Y=y\,|\,X=x)\}\\&=\iota\cdot\frac{\max\{E(d_X(X,x)\,|\,Y\neq y),E(d_X(X,x)\,|\,Y=y)\}-E(d_X(X,x))}{1-E(d_X(X,x))}\end{aligned}\tag{2}$$

where $\iota$ is $+1$ or $-1$ according to whether the average difference from $x$ is larger or smaller in the total population than among $y$-members [$\iota$ is the sign of $E(d_X(X,x))-E(d_X(X,x)\,|\,Y=y)$].

The absence of association of $y$ with $x$ is therefore characterized by equality of the overall and the conditional expectation, i.e. $E(d_X(X,x))=E(d_X(X,x)\,|\,Y=y)$. Obviously, stochastic independence between the trait $X$ and the state $y$ is sufficient for absence of association. It is however not sufficient to require stochastic independence only between the two states $x$ and $y$. On the other hand, there are special cases of stochastic dependence, where association is absent in terms of equality of the pertaining overall and conditional expectation.

## 5 State-to-Trait Level of Association

So far, association was regarded between individual states of two traits. The next higher level of association is that of the state of one trait with the entirety of states of another trait. At this level of association, one is interested in knowing whether possession of a particular state of one trait implies association with particular states of the other trait. Application of the approach taken at the state-to-state level suggests consideration of the difference between the $X$-states that $y$ is associated with and the $X$-states that $y$ is not associated with (or dissociated from). This corresponds to the idea that state $y$ can be distinguished or separated for its $X$-states from other $Y$-states (i.e. the remainder of the population). The more distinct this separation becomes, the stronger is the association of state $y$ with trait $X$. The association would be perfect if state $y$ is found to be associated with $X$-states with

which no other $Y$-state is associated, and if in addition the $X$-states with which $y$ is associated differ completely from the $X$-states in the remaining population. Recall that this perspective involves differences between $X$-states but not between $Y$-states. It therefore addresses the states of $X$ as potential *differentiae specificae* of the states of $Y$, where, for example, $y$ denotes a species that is distinguished from other species of the same genus by the states of trait $X$.

The problem to be addressed is apparently similar to that treated at the state-to-state level in that the separateness between $\Omega_y^c$ and $\Omega_y$ with respect to $X$-states determines the degree of association. The difference is that at the present state-to-trait level, separateness involves all states of trait $X$ rather than only one specific state. Hence, the previous concept of shift transformations applies identically to the assessment of separateness, where the separation is now determined by the differences between the $X$-states to which the individual shifts refer. A frequency shift $\sigma(a, b)$ from a source state $a$ to a sink state $b$ therefore entails a difference $d_X(a, b)$ between the two states. It follows that with each shift transformation $\sigma$ that matches the distribution of $X$ within $\Omega_y^c$ to that within $\Omega_y$, the pertaining total change in $X$-states amounts to

$$T_{y,X}(\sigma) = \sum_{a,b} \sigma(a,b) \cdot d_X(a,b) \tag{3}$$

$T_{y,X}(\sigma)$ can also be viewed to quantify the separation between $\Omega_y^c$ and $\Omega_y$ that goes along with the shift transformation $\sigma$.

The situation of complete separation and thus of perfect association of $y$ with $X$ is realized if for any pair $a$ and $b$ of $X$-states with $P(X = a \,|\, Y \neq y) > 0$ and $P(X = b \,|\, Y = y) > 0$ one has $d_X(a,b) = 1$. Since $d_X(a,a) = 0$, this implies that if either of the probabilities $P(X = a \,|\, Y \neq y)$ or $P(X = a \,|\, Y = y)$ is positive, then the other is zero. Hence, $\sigma(a,b) > 0$ only if $P(X = a \,|\, Y \neq y) > 0$ and $P(X = b \,|\, Y = y) > 0$ (which implies $d_X(a,b) = 1$). It follows that in this case

$$T_{y,X}(\sigma) = \sum_{a,b} \sigma(a,b) = 1 - \sum_{a} \min\{P(X = a \,|\, Y \neq y), \quad P(X = a \,|\, Y = y)\} = 1$$

so that $T_{y,X}(\sigma)$ reaches its maximum value of 1 only for perfect association of $y$ with $X$. On the other hand, there is no association if $X$ varies stochastically independently from $y$, and this shows as $\sigma(a,b) = 0$ for all $X$-states and therefore as $T_{y,X}(\sigma) = 0$. Apparently, $T_{y,X}(\sigma)$ displays features that are desirable for a measure of association.

However, it was mentioned earlier that there may be many shift transformations of one given distribution into another given distribution (for an illustration see Fig. 1). Since the objective consists in quantifying the separation between the distributions on the basis of state differences, it is essential to consider only shift transformations $\sigma$ that minimize the total change $T_{y,X}(\sigma)$. This suggests

$$\mathcal{A}(Y = y \,|\, X) := \min_{\sigma} T_{y,X}(\sigma) \tag{4}$$

as a *measure of association of state y with trait X*. The same approach was applied by Gregorius et al. (2003) to the measurement of differentiation between frequency distributions of traits with variable state differences. Algorithms for finding the

minimum separation (or differentiation) are described in this paper, and programs are available from the co-author E. Gillet.

In the case of perfect state-to-trait associations it is desirable to identify the involved $X$-states. By definition of the state-to-state associations, the $X$-states involved in $\mathcal{A}(Y = y \,|\, X) = 1$ are exactly those with which $y$ is perfectly associated, i.e. for which $\mathcal{A}(Y = y \,|\, X = x) = 1$. Conversely, if there are any states $x$ for which $\mathcal{A}(Y = y \,|\, X = x) = 1$ then $\mathcal{A}(Y = y \,|\, X) = 1$. Moreover, if $\mathcal{A}(Y = y \,|\, X = x) < 1$ for any $x$, then $y$ is perfectly dissociated from $x$ since the $X$-states not involved in $\mathcal{A}(Y = y \,|\, X) = 1$ differ completely from all $X$-states involved in $\mathcal{A}(Y = y \,|\, X) = 1$. This leads to the conclusion that $\mathcal{A}(Y = y \,|\, X) = 1$ if and only if for each $x$ the measure $\mathcal{A}(Y = y \,|\, X = x)$ equals either $+1$ or $-1$.

For highly variable $Y$-traits, but not only for these, it may happen that $y$-members are fixed for their $X$-trait. If the $Y$-trait is continuous this is even the normal case, since for such traits it is very unlikely that any two members of a population share a trait state. Thus typically $P(X = x \,|\, Y = y) = 1$ for some state $x$. Hence, there is only one sink state in $\Omega_y^c$, namely $x$, to which all other states are to be shifted. The only positive frequency shifts are therefore $\sigma(a, x) = P(X = a \,|\, Y \neq y)$ for $a \neq x$, so that Eq. 3 becomes $T_{y,X} = \sum_a P(X = a \,|\, Y \neq y) \cdot d_X(a, x) = \sum_a P(X = a) \cdot d_X(a, x) / P(Y \neq y)$ and consequently $\mathcal{A}(Y = y \,|\, X) = E(d_X(X, x) \,|\, Y \neq y)$ by Eq. 4. Hence, in this case the association of state $y$ with trait $X$ equals the average difference of $X$-states from $x$ among individuals not holding state $y$.

## 6 Trait-to-Trait Level of Association

At the next higher level of association (which is also the highest level) the state-to-trait level associations of the individual $Y$-states with trait $X$ are to be summarized. This summary should reflect the overall degree to which $Y$-states are distinguished for their $X$-states. The contributions of the individual $Y$-states correspond to their frequency, which specifies an appropriate *measure* $\mathcal{A}(Y \,|\, X)$ *of association of trait* $Y$ *with trait* $X$ by the averages of the state-to-trait level measures, i.e.

$$\mathcal{A}(Y \,|\, X) := \sum_y \mathcal{A}(Y = y \,|\, X) \cdot P(Y = y) \tag{5}$$

for trait $X$. In the absence of association $\mathcal{A}(Y \,|\, X) = 0$, so that there is no differentiation. Analogously, for perfect association, $\mathcal{A}(Y \,|\, X) = 1$, and differentiation is complete.

Perfect association can also be interpreted in terms of proper functional relations, since $\mathcal{A}(Y \,|\, X) = 1$ implies that $\mathcal{A}(Y = y \,|\, X) = 1$ for all $y$. As was shown above, this guarantees that members holding state $y$ do not share their $X$-states with members not holding $y$ and that between the two groups $X$-states differ completely. In other words, to each $X$-state there corresponds a unique $Y$-state, and the $X$-states corresponding to different $Y$-states are properly distinguished; thus $Y$ is a proper function of $X$. From this perspective, $\mathcal{A}(Y \,|\, X)$ presents itself as a measure of closeness of $Y$ to a proper functional dependence on $X$. Note that this distinguishes

$\mathcal{A}(Y \mid X)$ from the common indices of covariation (such as correlation or regression coefficients), which essentially measure deviations from models of linear (or more general monotonic) relationships between quantitative (or ordinal) variables. $\mathcal{A}$ applies to the detection of any functional relationship, including non-monotonic relationships as well as relationships between non-linear multi-dimensional or qualitative traits.

## 7 Association for Discrete Traits

Formally, discrete traits can be characterized by the existence of only two kinds of difference, one indicating sameness and the other differentness of states. Usually this amounts to choosing $d_X = 0$ or $d_X = 1$ according to whether the trait states of two individuals are identical or different, respectively. Hence, $E(d_X(X,x)) = P(X \neq x) = 1 - P(X = x)$, $E(d_X(X,x) \mid Y = y) = 1 - P(X = x \mid Y = y)$, and $E(d_X(X,x) \mid Y \neq y) = 1 - P(X = x \mid Y \neq y)$. At the state-to-state level of association, one therefore obtains from Eq. 2

$$\mathcal{A}(Y = y \mid X = x) = \iota \cdot \frac{P(X = x) - \min\{P(X = x \mid Y \neq y),\ P(X = x \mid Y = y)\}}{P(X = x)}$$

where $\iota$ is the sign of $P(X = x \mid Y = y) - P(X = x)$.

The state-to-trait level of association follows directly from Eq. 3 since

$$\begin{aligned} T_{y,X}(\sigma) = \sum_{a,b} \sigma(a,b) &= 1 - \sum_{a} \min\{P(X = a \mid Y \neq y),\ P(X = a \mid Y = y)\} \\ &= \frac{1}{2} \sum_{a} |P(X = a \mid Y \neq y) - P(X = a \mid Y = y)| \\ &= \mathcal{A}(Y = y \mid X) \end{aligned}$$

From this in turn the trait-to-trait level of association results as given in Eq. 5. All of these results conform with those derived earlier by Gregorius (1998). Also consult this paper for cases of further specialization to standard measures of association (such as measures of cross-classification, or linkage disequilibrium in population genetics).

## 8 Concluding Remarks

The areas of application of the above measures of association are quite diverse and cannot be appropriately represented in this paper. Therefore the following examples merely address opportunities for providing more detailed or alternative solutions to three problems of presumably common interest in population and ecological genetics. These will be preceded by a brief reference to functional relationships. In all examples the significance of considering both directions of association is pointed out.

• *Search for functional relationships*: At the state-to-state level, associations allow the creation of hypotheses about a functional relationship by consideration of both directions of association at the trait-to-trait level; choose the larger of the two associations to identify the direction of the functional relationship; for $\mathcal{A}(Y \mid X) > \mathcal{A}(X \mid Y)$ select for each $x$ a state $y$ such that $\mathcal{A}(Y = y \mid X = x) = \max_z \mathcal{A}(Y = z \mid X = x)$ and assign $y$ to $x$; this creates a hypothesis for a functional relationship $Y = f(X)$ (which need not be monotonic in any sense). Since the maximum association with $x$ may be realized by more than one state of $Y$, several hypotheses on a functional relationship $Y = f(X)$ may be possible.

For continuous variables, strict functional relationships may exist despite the fact that $Y$-states are not completely differentiated for their $X$-states. This is due to the condition that any statement on complete differentiation depends on difference measures for which complete difference is meaningfully defined. The latter, however, does not generally apply to continuous variables. Moreover, changes in $X$ may go along with small as well as with large changes in $Y$ (where the notions of "small" and "large" depend on the difference measure $d_Y$). The present concept of association therefore does not seem to cover such situations.

On the other hand, functional relationships, including continuous variables, are first of all based on the uniqueness of assignment of the states of the "independent" variable ($X$) to the states of the "dependent" variable ($Y$). Hence, only sameness or differentness of the states of the independent variable is relevant in this context, which in turn calls for a difference measure $d$ that allows unambiguous separation of all states. Given this, an additional binary difference measure $d^*$ can be specified by $d^* = 1$ if $d > 0$ and $d^* = 0$ if $d = 0$. As a result, $Y$ can be perfectly associated with $X$ when applying $d^*$, but the implied functional relationship may have to be considered imperfect when association is determined for the original difference measure $d$. The imperfection is thus due to unsatisfactory distinction or resolution of the states of trait $Y$ by the states of trait $X$.

• *Distribution of variation over populations*: A question frequently posed in population genetics concerns the mode according to which the genetic variation of a population is distributed over subpopulations. The same type of question is posed in ecology, where the distribution of the species in a region over communities is of concern. Both types of question can be tackled by characterizing each individual by its genetic type (or species affiliation) and by its subpopulation (or community) membership. Given that subpopulations are properly separated, subpopulation membership can be specified as a discrete trait with states that are equally different (binary difference measure). The latter does not apply to the genetic trait, since the genetic types may differ to variable degrees depending on the number of genes they share.

An assessment of the distribution of genetic variation over subpopulations can then be approached by computing the associations between the two traits (with $Y$ as subpopulation affiliation and $X$ as genetic type, for example). At the trait-to-trait level the association of $Y$ with $X$ would then be conceived as a measure of genetic differentiation among subpopulations. By consideration of genic differences between genotypes, it is possible to include gene interactions at different levels

as a new aspect into analyses of differentiation (Gillet and Gregorius 2008). The reverse association addresses differentiation between genetic types for their subpopulation memberships, as is relevant in the detection of tendencies of individuals with the same genotype to occur in the same subpopulation. This perspective is widely ignored in population genetic research (for an exception see Hudson 2000; a more comprehensive treatment can be found in Gregorius 2009).

• *Genomic associations*: Selection acting on pleiotropic or epistatically polygenic traits, inbreeding, or small population size entail evolutionary processes that shape in very different ways the multilocus genotypic structure of populations. These associations are commonly quantified in terms of linkage disequilibria or similar indices, all of which are based on haplotype frequencies (for an overview see e.g. Mueller 2004). For diploidy or higher degrees of ploidy, haplotype frequencies are difficult to estimate and may even miss important aspects of genomic association that show up at the genotypic level (for an example of linkage equilibrium with association at the genotypic level see Ziehe and Müller-Starck 1991, p. 260). Asymmetric associations seem to play no role at all in the context of linkage disequilibrium studies, even though they are to be expected at least among selected genetic traits or between selected and "background" traits.

The present measures of association apply to all degrees of ploidy, at least to the extent that appropriate measures of genic difference between genotypes are available (the total number of alleles by which two genotypes differ may be reasonable in many cases). This allows the study of associations between genomic regions that are chosen for their special functions and that need not be characterized by the same kind of genic difference measure. Another type of genetic association may exist between cellular organelles such as nuclear, mitochondrial, or plastid. These "cytonuclear disequilibria" (see e.g. Asmussen and Basten 1996) are defined in ways analogous to linkage disequilibria and can be substantially generalized with the help of the present measures.

• *Spatiogenetic structure*: This topic is frequently addressed in connection with dispersal problems and is analyzed in terms of spatial autocorrelation (for an overview see e.g. Epperson 2005). In essence, this type of analysis focuses on questions of covariation of genetic differences with spatial distances between pairs of individuals. In a more comprehensive (and probably intuitively more appealing) context, spatiogenetic structure can be conceived of as an association between genetic traits of individuals and their locations (as the second trait). The difference measures for the two traits can be specified by the genic difference between genotypes and by the spatial distance between individuals. Accordingly, association may be considered for genetic type with location and vice versa. Closeness to functional relationships of any shape (including monotonic relationships) can be assessed on the basis of the above demonstrations.

Association of genetic type with location targets situations where genetic types are differentiated for the locations in which they occur. Individuals differing in genetic type therefore tend to be separated more distinctly in space than individuals of the same genetic type. The degree to which the genetic types differ is not of concern so far. To allow for the assessment of closeness to perfect association, a

threshold distance is to be specified, beyond which spatial separation can be considered to be effectively complete. Thresholds are allowed for by the parameter $u$, which is applied in the normalization of difference measures to scales appropriate for association analyses (see Sect. 3). Perfect association is then reached if individuals of the same genetic type reside at the same location, while individuals with different genetic type are separated by a spatial distance of at least $u$.

If special threshold distances are not preset by the problem at hand, consideration of several thresholds is useful in order to identify distances for which the association changes distinctly. The present association measures allow for this kind of special inference, since they decrease (not necessarily strictly) with increasing $u$ (for a proof see "Appendix 1"). Thus, the thresholds could indicate the existence of distances within which family structures can build up, for example, while beyond the threshold distance individuals mix more or less freely. A more comprehensive picture of the distribution of genetic types in space therefore results from consideration of "association profiles", where the trait-to-trait level of association is plotted against the threshold values $u$. Association profiles can also be considered at the state-to-trait level in order to distinguish individual genetic types for their spatial distribution patterns.

In the reverse direction, association of locations with genetic types addresses differentiation among locations for the genetic types found at these locations. Apparently, spatial arrangement of the population members has no effect on association in this direction. If each location is occupied by a single individual only, the results of "Appendix 2" demonstrate that the association of locations with genetic types specializes to the average genetic difference between two different members of the population. Under the additional assumption of binary differences between genetic types, it turns out that association is identical to Simpson's index of diversity (Simpson 1949) when defined as the probability of drawing without replacement two individuals (locations) that differ in their genetic type (see "Appendix 2").

Locations may also be specified in terms of properly separated areas that are occupied by arbitrary numbers of individuals. Such areas may correspond to subpopulations, which, in accordance with the above perspective of the distribution of variation over populations, reveals the association of locations with genetic types to measure genetic differentiation among subpopulations.

There are of course many more opportunities for analyses of aspects of spatiogenetic structure inherent in difference measures and the corresponding association measures. One of these is provided by hierarchical clustering methods. In essence, a hierarchical clustering method transforms, via formation of its cophenetic differences, the initial difference measures into an ultrametric. Any ultrametric in turn represents a hierarchical (encaptic) structure (see e.e. Jardine and Sibson 1971, p. 50). Hence, application of a hierarchical clustering method to the locations and to the genetic types yields difference measures that are transformed such that they reflect various levels of spatial and genetic structure. This property is not affected by the demanded normalization of the differences. Since the ultrametric property implies equal difference between all members of two disjoint clusters, associations indeed refer to those of genetic types with spatial structure and to those of location with genetic structure.

**Acknowledgments** The author thanks the reviewers for their careful reading and commenting of the manuscript.



## Appendix 1

In the Sect. 3, different ways of normalizing difference measures are considered that are appropriate for the measurement of association. These normalizations are denoted by $\hat{d}$ and are functions of the initial (non-normalized) differences and the threshold parameter $u$. The normalizations fulfill the requirement that for any fixed difference value they decrease with increasing threshold parameter. The requirement is implied by the fact that complete difference, i.e. $\hat{d} = 1$, is reached only for initial differences equal to or greater than the threshold value.

As association is imperfect for differences smaller than the threshold difference, one expects that with increasing threshold level, association becomes even more imperfect, i.e. association measures decrease. More precisely, one expects that for any given non-normalized initial difference measure $d_X$, increasing the threshold value of the pertaining normalized difference measure $\hat{d}_X$ decreases the association $\mathcal{A}(Y = y \,|\, X)$. Indeed, this follows immediately from inspection of Eqs. 3 and 4 which specify the association $\mathcal{A}(Y = y \,|\, X)$ as resulting from a particular frequency shift transformation. Decreasing the differences $d_X$ (or $\hat{d}_X$, respectively) by increasing the threshold level obviously decreases $T_{y,X}(\sigma)$ for each shift transformation $\sigma$ and thus decreases $\mathcal{A}(Y = y \,|\, X)$. Hence, associations decrease at the state-to-trait as well as on the trait-to-trait level.

## Appendix 2

In the following, the effect of perfect association of trait $X$ with trait $Y$ on the reverse association of trait $Y$ with trait $X$ will be demonstrated. By definition, perfect association of $X$ with $Y$ entails that $X$ is a function of $Y$, i.e. $X = f(Y)$. This is equivalent to the statement that each state $y$ is fixed for its $X$-trait.

At the end of the section "state-to-trait level of association", it was shown that in this case $\mathcal{A}(Y = y \,|\, X) = E(d_X(X, f(y)) \,|\, Y \neq y)$, where $x$ is replaced by $f(y)$. Thus

$$\begin{aligned}\mathcal{A}(Y \,|\, X) &= \sum_y E(d_X(X, f(y)) \,|\, Y \neq y) \cdot P(Y = y) \\ &= \sum_y \sum_x d_X(x, f(y)) \cdot P(X = x \,|\, Y \neq y) \cdot P(Y = y)\end{aligned}$$

which equals the average difference between the $X$-states of two individuals that differ in their $Y$-state. This applies to the example in the Sect. 8, where $Y$ denotes the spatial position of an individual and $X$ is specified as a genetic trait.

If the measure of difference between $X$-states is binary, then perfect association of $X$ with $Y$ implies

$$\begin{aligned}\mathcal{A}(Y\,|\,X) &= \sum_y [1 - P(X = f(y)\,|\,Y \neq y)] \cdot P(Y = y) \\ &= 1 - \sum_y P(X = f(y)\,|\,Y \neq y) \cdot P(Y = y) \\ &= 1 - \sum_y [P(X = f(y)) - P(Y = y)] \cdot P(Y = y)/P(Y \neq y)\end{aligned}$$

The second equalities shows that this association equals the probability of drawing two individuals of different $Y$-state that differ in their $X$-state.

Assume in addition a uniform distribution of $Y$, so that $P(Y = y) = 1/N$ for all $y$, where $N$ denotes the number of $Y$-states. Then

$$\begin{aligned}\sum_y P(X = f(y)) &= \sum_x \sum_{y;f(y)=x} P(X = x) \\ &= \sum_x P(X = x) \cdot P(f(Y) = x) \cdot N \\ &= N \cdot \sum_x P(X = x)^2\end{aligned}$$

since $P(f(Y) = x) = P(X = x)$. Consequently, one obtains from the last of the above equalities for $\mathcal{A}(Y\,|\,X)$:

$$\begin{aligned}\mathcal{A}(Y\,|\,X) &= 1 - \frac{N^{-1}}{1 - N^{-1}} \cdot \left[ N \cdot \sum_x P(X = x)^2 - 1 \right] \\ &= \frac{N}{N - 1} \cdot \left[ 1 - \sum_x P(X = x)^2 \right]\end{aligned}$$

If each $Y$-state is represented by a single individual so that $N$ equals the population size, then the association is seen to be formally identical to Simpson's index of diversity of trait $X$ (defined as the probability of drawing without replacement two individuals that differ in their $X$-state).

## References

Asmussen MA, Basten CJ (1996) Constraints and normalized measures for cytonuclear disequilibria. Heredity 76:207–214

Epperson BK (2005) Estimating dispersal from short distance spatial autocorrelation. Heredity 95:7–15

Gillet EM, Gregorius H-R (2008) Measuring differentiation among populations at different levels of genetic integration. BMC Genet 9:60

Gillet EM, Gregorius H-R, Ziehe M (2004) May inclusion of trait differences in genetic cluster analysis alter our views? For Ecol Manage 197:149–158
Goodman LA, Kruskal WH (1979) Measures of association for cross classifications. Springer, New York
Gregorius H-R (1998) Measuring association between two traits. Acta Biotheor 46:89–98
Gregorius H-R (2009) Distribution of variation over populations. Theory Biosci 128:179–189
Gregorius H-R, Gillet EM, Ziehe M (2003) Measuring differences of trait distributions between populations. Biom J 45(8):1–15
Hudson RR (2000) A new statistic for detecting genetic differentiation. Genetics 155:2011–2014
Hughes TR (2008) 'Validation' in genome-scale research. J Biol 8:3. doi:10.1186/jbiol104
Jardine NJ, Sibson R (1971) Mathematical taxonomy. Wiley, London
Liebetrau AM (1983) Measures of association. Sage Publications, Beverly Hills
Liu L-Y (2005) Coefficient of intrinsic dependence: a new measure of association. Dissertation at the Texas A&M University, pp 76+xiii. http://hdl.handle.net/1969.1/2397
Hsing T, Liu L-Y, Brun M, Dougherty ER (2005) The coefficient of intrinsic dependence (feature selection using el CID). Pattern Recognit 38:623–636
Mueller JC (2004) Linkage disequilibrium for different scales and applications. Brief Bioinform 5:355–364
Simpson EH (1949) Measurement of diversity. Nature 163:688
Ziehe M, Müller-Starck G (1991) Changes of genetic variation due to associated selection. In: Müller-Starck G, Ziehe M (eds) Genetic variation in European populations of forest trees. J.D. Sauerländer's Verlag, Frankfurt am Main, pp 259–271