Theor Appl Genet (1987) 74:397–401

The relationship between the concepts of genetic diversity and differentiation

H.-R. Gregorius

Abteilung für Forstgenetik, Universität Göttingen, Büsgenweg 2, D-3400 Göttingen, Federal Republic of Germany

Received January 26, 1987; Accepted February 23, 1987
Communicated by P. M. A. Tigerstedt

Summary. Diversity as a measure of individual variation within a population is widely agreed to reflect the number of different types in the population, taking into account their frequencies. In contrast, differentiation measures variation between two or more populations, demes or subpopulations. As such, it is based on the relative frequencies of types within these subpopulations and, ideally, measures the average distance of subpopulations from their respective lumped remainders. This concept of subpopulation differentiation can be applied consistently to a single population by regarding each individual as a deme (subpopulation) of its own, and it results in a measure of population differentiation δ_T which depends on the relative frequencies of the types and the population size. δ_T corresponds to several indices of variation frequently applied in population genetics and ecology, and it verifies these indices as measures of differentiation rather than diversity. For any particular frequency distribution of types, the diversity v is then shown to be the size of a hypothetical population in which each type is represented exactly once, i.e. for which $\delta_T = 1$. Hence, the diversity of a population is its differentiation effective number of types. This uniquely specifies the link between the two concepts. Moreover, v again corresponds to known measures of diversity applied in population genetics and ecology. While population differentiation can always be estimated from samples, the diversity of a population, particularly if it is large, may not be. In such cases, it is recommended that population differentiation is estimated and the corresponding sample diversity merely computed. Finally, a solution to the problem of measuring multi-locus diversities is provided.

Key words: Genetic diversity – Population differentiation – Multi-locus diversity

The problem

Although diversity is one of the most fundamental concepts in experimental and theoretical ecology, population biology and population genetics, many different measures of diversity are still in use. This is particulary unsatisfactory, since even the ranking of communities or populations for diversity may depend on the measure used, as was demonstrated by Ziehe (1982). Moreover, some of the measures are based on reasoning derived from specific models or probabilistic concepts which show only a loose connection with the general idea of diversity. In some cases it is not even quite clear whether these measures might be more suitable for the quantification of concepts other than diversity. It appears to be widely accepted that diversity ought to reflect the number of different types in a collection of objects (community, population etc.), where the types are determined by the expressions of a trait under consideration. Ideally, if all types are equally frequent, diversity should directly relate to the number of types, and diversity should decrease with increasing deviation from a uniform distribution of types. The more dominant in frequency one type becomes, the closer a diversity measure should come to its minimun value representing the absence of diversity, i.e. the presence of only one type (monomorphic collections).

Supplementing these basic requirements by a few intuitively reasonable conditions, Routledge (1979) and Gregorius (1978) independently came to the conclusion that a diversity measure v should be of the form

$$v_a = \left(\sum_{i=1}^{n} p_i^a \right)^{1/(1-a)},$$

where p_i is the relative frequency (probability) of i-type objects in the collection and a is a positive real

number not equal to 1. The latter author proved that as a approaches a value of 1, v_a approaches

$$v_1 = p_1^{-p_1} \cdot p_2^{-p_2} \cdot \ldots \cdot p_n^{-p_n},$$

the logarithm of which is the well known information or Shannon-Wiener measure of diversity. Similarly, as a approaches infinity, v_a becomes

$$v_\infty = (\max p_i)^{-1}.$$

This is particularly noteworthy since the authors had different objects in mind, the first referring to community and the second to genetic diversity. The class of v_a-measures has the property that for a given set of non-uniformly distributed frequencies, diversity decreases strictly with increasing a. Irrespective of the value for a, v_a becomes equal to the number of types if these are uniformly distributed in the collection. Thus the minimum value for v_a is 1, and this corresponds to a monomorphic collection. The authors also agreed that $a=2$ might be the best choice, although for less objective reasons.

A second concept which is frequently applied, particularly in population genetics, is that of differentiation. It is designed to measure the amount of genetic differences between parts (demes, subpopulations, etc.) of a population, and it is based on relative frequencies of genes. Hence, although both concepts are concerned with the measurement of variation, they differ in that diversity applies to variation within a single collection and differentiation applies to variation between an arbitrary number of collections of objects. The most frequently used measure of genetical differentiation among subpopulations is Wright's F_{ST} (1978, chapter 3), which is identical to Nei's G_{ST} (1973). F_{ST} ranges between 0 and 1; it becomes 0 if all subpopulations are genetically identical (have the same gene frequencies) and it is 1 only if all subpopulations are genetically monomorphic. As Wright himself noticed, the latter property is a serious weakness of F_{ST}, since it classifies a population as completely differentiated even in cases where some, but not all, of its subpopulations are fixed for the same allele.

For this reason, Gregorius and Roberds (1986) suggested a new measure δ of subpopulation differentiation, which is unique for a fairly small number of conceptually cogent conditions (this paper also provides a more detailed discussion of F_{ST} as opposed to δ). Among these are that $\delta=0$ only if all demes are genetically identical and that $\delta=1$ only if all demes are genetically unique ($0 \leqq \delta \leqq 1$). Thus, $\delta=0$ indicates the absence of differentiation (uniformity), and $\delta=1$ indicates complete differentiation with respect to the population subdivision considered. The principle consists in measuring for each deme its genetic distance from the respective remainder of the population and taking the weighted average over all demes, where the weights are given by the deme sizes. The appropriate measure of genetic distance between two populations p and q, say, was shown to be

$$d_0(p,q) = 0.5 \cdot \sum_i |p_i - q_i|,$$

where p_i and q_i are the relative frequencies of the i-th genetic type in population p and q, respectively (note that the p's and q's are not restricted to gene frequencies but may also be applied to gametic, single- and multi-locus genotypic frequencies, etc.). With this representation, δ measures the proportion of the total number of genetic elements in the population by which the demes effectively differ from their respective complements.

In an attempt to analyse genetic variation between and within demes of a subdivided population, Nei (1973) used the notions of diversity and differentiation interchangeably. For example, following his equation (7) he terms his D_{ST} "the average gene diversity between subpopulations". Three sentences later he states that the "absolute magnitude of gene differentiation among subpopulations may be measured by D_{ST}" and proceeds to define "gene differentiation relative to the total population" as $G_{ST} = D_{ST}/H_T\,(=F_{ST})$, where H_T is termed the "gene diversity in the total population". Nei didn't provide arguments in support of his terminology other than those immediately emanating from the particular decomposition of sums of squares of gene frequencies considered. In the light of the basic features of the two concepts explained above, this is somewhat confusing, since the measurement of genetic variation *within* subpopulations (diversity) appears to be confounded with that of variation *between* subpopulations (differentiation). However, it also indicates that there might be an intrinsic relationship between the two concepts which we might not always be fully aware of.

In the following is shall be briefly demonstrated that this relationship can be obtained by consistently extending the notion of differentiation between populations to differentiation within a single population. Using δ as the basic measure, it will be shown that diversity and differentiation have a common origin lying in the genetic distance d_0. It will turn out that diversity (v) can be simply conceived of as the 'differentiation effective number of types', and that indeed $v = v_a$ with $a=2$. Although the primary concern of this paper is of a genetical nature, most of the results will be easily seen to apply to more general situations.

Differentiation

Consider a population which is subdivided into demes (subpopulations) of relative size c_j (for the j-th deme,

$\sum_j c_j = 1$). Let the relative frequency of genetic elements (or any other objects) of type i in deme j be $p_i(j)$, $\sum_i p_i(j) = 1$. Moreover, let $\bar{p}_i(j)$ be the relative frequency of i-types in the complement of the j-th deme, i.e. $\bar{p}_i(j) = \sum_{k, k \neq j} p_i(k) \cdot c_k/(1-c_j)$. Then, according to Gregorius and Roberds (1986), the amount of subpopulation differentiation with respect to the given subdivision can be measured by

$$\delta = \sum_j c_j \cdot d_0(p(j), \bar{p}(j)),$$

where $d_0(p(j), \bar{p}(j)) = 0.5 \cdot \sum_i |p_i(j) - \bar{p}_i(j)|$ is the genetic distance between the j-th deme and its complement.

Now, consider the special case in which each individual genetical element is regarded as a deme of its own, and suppose that there are N_i genetic elements of type i. Then, $N = \sum_i N_i$ represents the population size, and, by assumption, each deme has relative size $1/N$. The d_0-distance between an individual genetic element of type k and its population complement is

$$0.5 \cdot \left(|1-(N_k-1)/(N-1)| + \sum_{i, i \neq k} |0-N_i/(N-1)| \right) = \frac{N-N_k}{N-1}.$$

Consequently,

$$\delta = \sum_k \frac{N_k}{N} \cdot \frac{N-N_k}{N-1}.$$

Denoting by $p_i = N_i/N$ the relative frequency of i-types in the population and setting $\delta = \delta_T$ (to distinguish it from ordinary subpopulation differentiation),

$$\delta_T = \frac{N}{N-1} \cdot \left(1 - \sum_i p_i^2\right)$$

can consistently be termed the degree or level of *population differentiation* since it refers to the total population by considering each deme to consist of exactly one individual genetic element. Thus δ_T is the proportion of the total number of genetic elements in the population by which the individual genetic elements effectively differ from their respective complements (cf. Gregorius and Roberds 1986). Moreover, the way δ_T is derived shows that the concept of differentiation applies to the measurement of variation between demes as well as within populations, and in the latter case it is, of course, independent of subpopulation structure.

However, δ_T also has a probabilistic interpretation: it is identical to the probability that two individuals (individual genetic elements) sampled at random, and without replacement, differ in type. With this interpretation δ_T is identical to what is called Simpson's measure of 'diversity' (Simpson 1949; see also Pielou 1969, p. 223). Yet, as we now see 'diversity' is not the appropriate characterization of this probability. This is also implicit in Pielou's (1969, p. 103) statement that "it is reasonable to call the diversity of a collection great if this probability ($1-\delta_T$) is low and slight if it is high". Hence, she implies that $1-\delta_T$ and thus δ_T is only indirectly related to but not by itself a measure of diversity. The exact relationship will be provided in the next section.

For effectively infinite population size, $\delta_T = 1 - \sum_i p_i^2$. Under this assumption the sample differentiation

$$\delta_T^* = \sum_i \frac{N_i}{N} \cdot \frac{N-N_i}{N-1},$$

where the N_i's are now the sample frequencies of i-types and N is the sample size, is a consistent and unbiased estimator of δ_T. Thus, if the p_i's are gene frequencies, Nei's (1973) measure of 'gene diversity' $H = 1 - \sum_i p_i^2$ actually measures gene differentiation in an effectively infinite population, and an appropriate estimator would be specified by δ_T^*.

Diversity

As was already pointed out, the notion of diversity basically refers to the absolute number of different types in a collection or population, and there are measures (the v_a's) which account for this, including the particularly relevant situation in which the types are unevenly distributed. The latter implies that diversities cannot only be natural numbers. This immediately becomes clear if one considers the example where a population consists of only two types, one of which dominates the other in frequency. In this case the number of types is effectively neither one nor two, but rather some value between these numbers. Hence, in some sense diversity traces the actual population back to an ideal one, in which all types are equally represented, but not neccessarily in terms of natural numbers. Yet, natural numbers set the frame for interpretation, so that a diversity of 7.8 would mean that the collection effectively contains seven to eight types, lying closer to 8. The problem with the existing measures of diversity (including all v_a's) is that, for the same population, they produce values differing by more than 1 (see figures, Gregorius 1978), which makes an interpretation in terms of the effective number of types difficult as long as a reference is not explicitly specified.

The previous considerations suggest that the degree of population differentiation δ_T be used as a reference. In complete differentiation, where each type is represented only once in the population ($\delta_T = 1$), the diversity must be identical to the population size N. Thus, for any particular frequency distribution of types (p_i's) we can ask for the size of a hypothetical population, in which each type is represented exactly once, i.e. for which $\delta_T = 1$. In other words, we seek a hypothetical size v for which

$$\frac{v}{v-1} \cdot \left(1 - \sum_i p_i^2\right) = 1 .$$

This v can then consistently be defined as the diversity of a population corresponding to its degree of differentiation and shall be called the *differentiation effective number of types.* It follows that

$$v = \left(\sum_i p_i^2\right)^{-1},$$

which is identical to one of the v_a-measures of diversity, namely that with $a = 2$, and is thus in agreement with the recommendations of Routledge (1979) and Gregorius (1978). When applied to allele frequencies, v is also identical to the 'effective number of alleles' of Crow and Kimura (1970, p. 323). Moreover, the general relationship between population diversity and differentiation can now be written in the form

$$\delta_T = \left(1 - \frac{1}{v}\right) \cdot \frac{N}{N-1} = \frac{1 - \frac{1}{v}}{1 - \frac{1}{N}},$$

which becomes $\delta_T = 1 - 1/v$ for large population size. This is a pleasing result in that it interrelates the two concepts by tracing them back to the basic problem of measuring differences, the solution of which is provided by the distance measure d_0.

At first sight, the approach of Rao (1982) may appear similar to the present approach. Aiming at a unifying representation of what he calls diversity and dissimilarity coefficients, Rao chooses general difference functions between characteristics of individuals as the common basis and arrives at his coefficients by averaging these differences within and between (pairs of) populations. The latter accounts for the basic conceptual difference between diversity and differentiation (Rao's dissimilarity). However, the author does not specify his ideas of diversity and dissimilarity, so that it is not clear whether his individual difference functions and their averages can be appropriate measures. Moreover, diversity appears to be a particular case of dissimilarity in Rao's paper, which blurs the intrinsic difference between the two concepts.

Differentiation effective diversity at multiple loci

Clearly, if multi-locus genotypes or gametes can be determined, diversity can be directly computed with the help of the previously given formula. A problem arises if allelic frequencies can be obtained only separately for several loci, as was pointed out by Chambers and Bayless (1983). Gregorius and Roberds (1986) demonstrate that, in this case, the gene pool has to be considered as the underlying collection of objects. The authors derived the gene pool differentiation as the average taken over the single locus levels of differentiation. Consistent extension of this result to the present measure δ_T of population differentiation leads us to

$$\delta_T = \frac{1}{L} \cdot \sum_{l=1}^{L} \delta_{T(l)} ,$$

where L is the number of gene loci and $\delta_{T(l)}$ is the population differentiation at the l-th locus. Let $p_{i,l}$ be the relative frequency of the i-th allele at the l-th locus $\left(\sum_i p_{i,l} = 1\right)$, and let N' be the population size, so that there are $N = 2 \cdot N'$ individual alleles at each diploid locus. Then

$$\delta_T = \frac{N}{N-1} \cdot \frac{1}{L} \cdot \sum_{l=1}^{L} \left(1 - \sum_i p_{i,l}^2\right),$$

and, applying the previous findings, the differentiation effective number of alleles per locus (the per locus gene pool diversity) is obtained by setting $\delta_T = 1$ in the above equation and solving for N. Denoting by

$$v_{(l)} = \left(\sum_i p_{i,l}^2\right)^{-1}$$

the allelic diversity at the l-th locus, the *gene pool diversity* v turns out to be

$$v = \left(\frac{1}{L} \cdot \sum_{l=1}^{L} \frac{1}{v_{(l)}}\right)^{-1} .$$

Thus, the gene pool diversity is equal to the harmonic mean of the single locus diversities, while the gene pool differentiation is equal to the arithmetic mean of the single locus differentiations.

Conclusions

It has now become clear that, while the (differentiation effective) diversity v of a population depends on the relative frequencies of the types only, the population differentiation δ_T additionally depends on the population size N, and $0 \leqq \delta_T \leqq 1 \leqq v \leqq N$. For a given v, δ_T is a hyperbolically decreasing function of N, so that the influence of population size on the degree of popula-

tion differentiation vanishes rather rapidly with increasing N. For example, if $N \geqq 30$, then δ_T differs from its limiting value at infinite population size by a factor of less than 1.04. This might lead us to conclude that in most cases of practical relevance, δ_T and v are mere one-to-one transformations of each other, and that, therefore, the decision to use one of these two measures depends solely on whether one likes values smaller or greater than 1. There is at least one important reason why, particularly in large populations, the two concepts should not be thought to make equivalent statements. Differentiation provides information relative to the whole population, while diversity relates to an absolute number. Suppose, for example, that a very large population is completely differentiated with respect to some trait, i.e. each individual is different in type from every other. This may easily happen if the genotype of individuals in a sexually reproducing population can be determined for a large number of polymorphic loci. For practical reasons, estimates of population size may not be obtainable, and all information on the amount of variation has to be extracted from comparatively small samples. Then, in each sample of size N the diversity v would be $v = N$, and diversity estimates would be completely correlated with sample size, i.e. increasing the sample size would automatically increase the diversity estimate. Thus, the problem of estimating diversity would be insoluable. In contrast, the sample differentiation, which, as was previously noted, is a consistent and unbiased estimator of population differentiation, would immediately teach us that the population is completely differentiated at a specified level of significance and irrespective of the actual population size.

Since this also applies to less extreme conditions, the concept of differentiation might be more suitable and effective for the assessment of the amount of variation than the concept of diversity. In addition, the fact that diversity measures an absolute population value makes it impossible to evaluate its proper significance without knowing the population size N simply because N is the upper limit for diversity. Thus, for unknown N we are not able to conclude from a v-estimate whether a population is very diverse or not. Differentiation estimates are not subject to such criticism, because by definition they are relative measurements and thus do not depend on N.

This is, of course, not to say that the concept of diversity is generally irrelevant. If the number of types found in a sample is either distinctly smaller than the sample size or if the types are unevenly distributed, the sample diversity might still provide quite reliable information about the effective number of types in the population. Moreover, there are many situations conceivable for which absolute numbers of types are required, even if they can be determined only for the sample itself.

Acknowledgements. The author appreciates the comments of H. H. Hattemer, G. Müller-Starck, K. Radler, J. Roberds and M. Ziehe.

References

Chambers SM, Bayless JW (1983) Systematics, conservation, and the measurement of genetic diversity. In: Schonewald-Cox CM, Chambers ST, MacBride B, Thomas L (eds) Genetics and conservation. Benjamin/Cummings, London Amsterdam; Don Mills, Ontario Sydney Tokyo, pp 349–363

Crow JF, Kimura M (1970) An introduction to population genetics theory. Harper & Row, New York Evanston London

Gregorius H-R (1978) The concept of genetic diversity and its formal relationship to heterozygosity and genetic distance. Math Biosci 41:253–271

Gregorius H-R, Roberds JH (1986) Measurement of genetical differentiation among subpopulations. Theor Appl Genet 71:826–834

Nei M (1973) Analysis of gene diversity in subdivided populations. Proc Natl Acad Sci USA 70:3321–3323

Pielou EC (1969) An Introduction to Mathematical Ecology. Wiley & Sons, New York London Sydney Toronto

Rao CR (1982) Diversity and dissimilarity coefficients: a unified approach. Theor Popul Biol 21:24–43

Routledge RD (1979) Diversity indices: which ones are admissible? J Theor Biol 76:503–515

Simpson EH (1949) Measurement of diversity. Nature 163:688

Wright S (1978) Evolution and the genetics of populations, vol 2. University of Chicago Press, Chicago

Ziehe M (1982) Quantifizierung genetischer Variation. Forum Genetik-Wald-Forstwirtschaft. Bericht über die 2. Arbeitstagung, Göttingen, S 41–49