Measure of Multilocus Correlation as a New Parameter for Study of Plant Pathogen Populations

E. Kosman

Institute for Cereal Crops Improvement (ICCI), The George S. Wise Faculty for Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel.
Accepted for publication 27 May 2003.

ABSTRACT

Kosman, E. 2003. Measure of multilocus correlation as a new parameter for study of plant pathogen populations. *Phytopathology* 93:1464-1470.

The measure of multilocus correlation for diallelic loci was developed to augment use of other diversity indices in the study of multilocus structure of populations. This measure provides information not revealed by other parameters for measuring multilocus association and linkage disequilibrium among pairs of loci. Relationships between the measures of multilocus correlation and association and all commonly used indices for diversity within populations are also described. A number of hypothetical examples demonstrate that the measure of multilocus correlation describes unique aspects of multilocus structure of populations compared with the measure of multilocus association and ϕ coefficient of associ-

ation. The measure of multilocus correlation can be used for analysis of independence of differentiating characters for a given population—the necessary condition of valid applications of the bootstrap method across differentials. Comparisons of the measures of multilocus correlation and association with the measures of diversity within population show that they reveal different aspects of population structure. As an adjunct to standard diversity indices, the measure of multilocus correlation may reveal subtle differences in diversity within populations even if the standard index fails to distinguish between the populations. A new measure of population uniformity also was developed to characterize another aspect of diversity within populations not detected by standard indices.

Additional keywords: Kosman diversity index, population genetics.

Study of diversity within plant pathogen populations is usually based on different biological methods of analyses such as virulence typing or comparison of isozymes or DNA markers by random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), and restriction fragment length polymorphism (RFLP). The formal structure of data is the same for all these methods. A number of isolates are tested on a set of differentials, which can be single-gene isolines in the case of virulence typing or different DNA bands in the cases of RAPD, AFLP, and RFLP analyses. The multilocus patterns produced by the tested isolates are tabulated in a two-way data table that consists of 0s and 1s in the case of two alleles for all loci or bands. Several methods (9,12,13,15,17) for measuring diversity within plant pathogen populations are based on such data. However, the standard methods do not help to reveal possible relations between differentiating characters (loci) or demonstrate how these relationships contribute to the population diversity.

The concept of linkage or association of alleles among different loci is central in population genetics (4,20). Gametic disequilibrium within populations of plant pathogens has been studied by many researchers (2,3,7,8,10,14,21,22). The Fisher's exact test (19) and its improved version (18) for contingency tables has usually been used for measuring association (disequilibrium) between pairs of the virulence and molecular (RAPD and RFLP) markers (7,8,10), whereas the multilocus association (2,3,14,22) has usually been tested using the method developed by Brown et al. (4). However, the contribution of differentials' association for a pathogen population to the degree of diversity within population has not been explored adequately.

Independence/dependence of differential factors is an important characteristic of the multilocus structure of any population. In

particular, "independence" of differential factors is required for correct applications of the bootstrap method by resampling across differentials (5,6). If dendrograms obtained by the bootstrapping across differentials are used to determine confidence levels in cluster analysis, the differentiating characters need to be more or less independent for the validity of the conclusions relative to confidence levels assigned to branches in the resultant dendrogram. As will be demonstrated, some differential characters used for cluster analyses may have "correlated effects" even when standard tests for linkage disequilibrium or multilocus association do not reveal any association. However, even the assumption of no strong association between characters is not usually checked. The "independence" between characters means no strong effect of correlation. Many plant pathogen populations often have high correlations among virulence and molecular markers because they reproduce asexually. The conclusions from bootstrap analyses will be distorted if the differential factors used to characterize the isolates have strong correlated effects. Until now, no method has been proposed to measure such correlated effects between characters that may not be revealed in tests of association or to measure overall correlation between differentiating parameters (multilocus correlation). Furthermore, the multilocus correlation can provide useful information on diversity in comparisons between pathogen populations tested on the same set of differentials and may reveal subtle differences that would not be apparent in comparisons of standard diversity indices.

The objectives of this study were to develop a measure of multilocus correlation (overall correlation between differentiating characters), to compare it with standard diversity indices and the measure of multilocus association, and to show how the measure of multilocus correlation may be used in combination with appropriate indices in analyzing genetic diversity in populations of plant pathogens. In particular, we will demonstrate that the measure of multilocus correlation restricted for two differentiating characters may reveal strong correlated effects between them and can be used for valid implementation of the bootstrapping across charac-

Corresponding author: E. Kosman; E-mail address: kosman@post.tau.ac.il

Publication no. P-2003-1007-01O
© 2003 The American Phytopathological Society

ters. The comparisons and explanations are made in terms of virulence typing data but they can easily be extended for other kinds of data. Terminology similar to that of statistics is used, but the terms “independence” and “correlation” are not necessarily identical to the corresponding statistical notions.

MEASURE OF MULTILOCUS CORRELATION

Let us start with the simplest data when n isolates of a plant pathogen are tested for virulence/avirulence on the set of differentials D_1 and D_2 . The distance, $d(D_1, D_2)$, between two differentials is defined as the number of isolates with different virulence type on the given differentials. If each isolate has the same virulence type, then the distance between the differentials equals 0. If each isolate does not share the same virulence type on both differentials, then the distance between the differentials equals n . These two cases are extreme: the first one is characterized by the absolutely positive correlation between the differentials, whereas the second case is characterized by the absolutely negative correlation. It is natural to assume that a measure of correlation between any two differentials has to be of the greatest and the smallest value in the first and the second case, respectively. Therefore, we define the measure of correlation between two differentials by

$$Cor(D_1, D_2) = 1 - 2 \cdot \frac{d(D_1, D_2)}{n} \quad (1)$$

This is the normalized measure, which ranges from -1 to 1 : $Cor(D_1, D_2) = 1$ if each isolate has the same virulence type, and $Cor(D_1, D_2) = -1$ if each isolate does not share the same virulence type on both differentials. Differentiating characters are strongly correlated if $Cor(D_1, D_2)$ is relatively close to -1 or to 1 , whereas there is no correlation between them if the values of $Cor(D_1, D_2)$ are close to zero.

Two characters are considered absolutely independent if the measure of correlation between them equals zero. Any deviation from the absolute independence can be interpreted as evidence of dependence between differentiating characters. Choice of a suitable threshold may be subjective and clearly justified by the character and goals of a specific study. Two characters are absolutely dependent if the measure of correlation between them equals ± 1 .

Let p_i and $1 - p_i$ denote frequencies of two alleles (designated 1 and 0, respectively) at diallelic locus i ($i = 1, 2$). In the case of virulence data, p_i and $1 - p_i$ are frequencies of virulent and avirulent isolates, respectively, on differential D_i , $i = 1, 2$. Four different patterns of individuals are possible with respect to these two loci: 11, 10, 01, and 00. We denote frequencies of these phenotypes for any sample of isolates as q_{11} , q_{10} , q_{01} , and q_{00} , respectively. Then correlation between two differentials also can be expressed by

$$Cor(D_1, D_2) = 1 - 2 \cdot (q_{10} + q_{01}) \quad (1a)$$

The following inequalities are fulfilled

$$1 - 2 \cdot \min\{2 - (p_1 + p_2), p_1 + p_2\} \leq Cor(D_1, D_2) \leq 1 - 2 \cdot |p_1 - p_2| \quad (1b)$$

where $\min\{a, b\}$ means minimum of the two values a and b . These inequalities determine the range for values of correlation between two differentials for given values of virulence frequencies p_1 and p_2 . The exact value of correlation belongs to this interval and depends on the degree of association between the two differentials.

Let us study a general case when the data obtained from testing the virulence of a plant pathogen to a set of k differentials is considered. Formally, an arbitrary sample from a pathogen population, A , is presented by n patterns of dichotomous responses of its isolates to some set D of k factors, $D = \{D_1, D_2, \dots, D_k\}$. The dis-

tance, $d(D_i, D_j)$, between two factors (differentials) i and j equals the number of isolates that respond differently to D_i and D_j ($i, j = 1, 2, \dots, k$). To each differential, match a differential from the set D to make up k pairs to maximize the sum of distances between corresponding pairs of differentials. Finding such matches can be realized by the solution of the appropriate assignment problem (1,9). If $M(D)$ is the obtained maximum value of the sum of distances between matched pairs of differentials, then the measure of overall correlation between differentials from the set D with respect to the given sample from the pathogen population A is defined by

$$Cor_A(D) = 1 - 2 \cdot \frac{M(D)}{n \cdot k} \quad (2)$$

where k is the number of differentiating factors and n is the number of sampled isolates. We also call this measure multilocus correlation. If a set of differentials is fixed, then the measure of multilocus correlation $Cor_A(D)$ is denoted as $Cor(A)$.

Formula 1 is a partial case of formula 2. Thus, for a set of two differentials only, $D = \{D_1, D_2\}$, we have $M(D) = d(D_1, D_2) + d(D_2, D_1) = 2d(D_1, D_2)$ and $k = 2$.

Example 1. Consider the response patterns of six isolates ($n = 6$) from population A to the set D of four differentials ($k = 4$) D_1 , D_2 , D_3 , and D_4 , respectively,

	1	2	3	4
A	1	1	1	1
	1	0	1	0
	1	0	0	0
	0	0	0	1
	0	0	1	1
	0	0	1	1

The distances between all paired differentials are determined as follows:

d	D_1	D_2	D_3	D_4
D_1	0	2	3	5
D_2	2	0	3	3
D_3	3	3	0	2
D_4	5	3	2	0

In general, there are 24 ($k!$ for $k = 4$) possibilities to match four differentials. The solution $M(D)$ of the assignment problem (matching differentials for maximum) is obtained for the following set of four pairs of matched differentials: (D_1, D_4) , (D_2, D_3) , (D_3, D_2) , and (D_4, D_1) . This maximum is $M(D) = d(D_1, D_4) + d(D_4, D_1) + d(D_2, D_3) + d(D_3, D_2) = 5 + 5 + 3 + 3 = 16$. (Generally, another set of matched differentials also may provide the solution of the assignment problem, but this maximum $M(D)$ is the same unique value for each possible set and characterizes the given sample of isolates.) Then, according to formula 2, the value of multilocus correlation is

$$Cor_A(D) = 1 - 2 \cdot \frac{16}{6 \cdot 4} = -\frac{1}{3}$$

The measure of multilocus correlation ranges from -1 to 1 . The only case in which the overall correlation score is assigned 1 corresponds to a population with identical response of each isolate on the entire set of differentiating factors, i.e., each isolate is virulent or avirulent to all differentials and, therefore, $d(D_i, D_j) = 0$ for all $i, j = 1, 2, \dots, k$. The minimal multilocus correlation score is assigned to a population that is tested on a few pairs of differentials with an absolutely different virulence type of all isolates on differentials of each pair from the entire set of pairs of differentiating factors, i.e., $d(D_i, D_j) = n$ if D_i and D_j form any mentioned pair of differentials. If values of multilocus correlation are close to 1, then the corresponding sets of differentials necessarily include pairs of differentials on which nearly all individuals respond identically

(pairs of mutual positive relationship). If values of multilocus correlation are close to -1 , then the corresponding sets of differentials necessarily include pairs of differentials on which nearly all individuals respond differently (pairs of mutual negative relationship). In all other cases, the corresponding sets of differentiating characters may include both positively and negatively correlated pairs.

The measure of multilocus correlation is a global parameter for the given set of differentials D . Let DS be any subset of differentials from D . It is easy to show that $Cor_A(D) = 1$ if and only if $Cor_A(DS) = 1$ for each arbitrary subset DS . But almost nothing else can be stated about $Cor_A(DS)$ and $Cor_A(D)$ if one of these two measures is known. Thus, it can be possible to select subsets of differentials for which the multilocus correlation differs from that of the parent set.

Example 2. Consider two populations A and B tested on set D of eight differentials and presented by the following data:

Population A								Population B							
1	2	3	4	5	6	7	8	1	2	3	4	5	6	7	8
1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	0
1	1	1	1	0	0	0	0	1	1	1	1	1	1	1	0
0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	1
0	0	0	0	1	1	1	1	0	0	0	0	0	0	0	1

The overall correlation $Cor_A(D) = -1$ is minimal, but there is subset DS comprising differentials D_1, D_2, D_3 , and D_4 with the maximal value of the overall correlation, $Cor_A(DS) = 1$. On the other hand, the overall correlation between differentials from D with respect to population B is close to 1, $Cor_B(D) = 0.8$, whereas $Cor_B(D_7, D_8) = -1$ for the set of two differentials D_7 and D_8 .

MEASURES OF MULTILOCUS CORRELATION AND ASSOCIATIONS

We consider two measures of association: the usually used ϕ coefficient of association between pairs of differentiating characters (equation 17.5 in literature citation 19) which is related to Fisher's exact test (19), and the measure of multilocus association, Ass , developed by Brown et al. (4). The measures of correlation (Cor) and association (ϕ and Ass) are aimed at characterizing a plant pathogen population by study of relationships between differentiating parameters. These measures appear somewhat similar but they describe different features of differential sets. On the one hand, if allele frequencies are different from 0 and 1, the extreme values of correlation (1 and -1) between two loci result in the extreme values (1 and -1 , respectively) of the ϕ coefficient of association and vice versa, which reflects equivalence of the absolute correlation and the absolute linkage disequilibrium among two loci in this case. On the other hand, a strong correlation also may be observed even in the case of absolute gametic equilibrium, and in contrast, two differentials can be independent ($Cor = 0$) despite association among them. The following examples demonstrate that correlation and linkage of alleles among different loci are separate characteristics and the relationships between them may not be predictable.

Example 3. If frequencies of alleles are of either very low or very high levels in the case of two diallelic loci, then the measure of correlation between these loci are of relatively high levels. This follows from inequalities 1b if virulence frequencies p_1 and p_2 are close to either 0 or 1.

Example 3.1. Let us consider a sample of isolates tested on two differentials, D_1 and D_2 , with virulence frequencies $p_1 = p_2 = 0.05$. Frequencies of four possible phenotypes 11, 10, 01, and 00 equal $q_{11} = p_1 \cdot p_2 = 0.0025$, $q_{10} = p_1(1 - p_2) = 0.0475$, $q_{01} = (1 - p_1)p_2 = 0.0475$, and $q_{00} = (1 - p_1)(1 - p_2) = 0.9025$, respectively, in the case of random distribution of virulences among isolates. Therefore, there is no linkage between the differentials, which is confirmed by the ϕ coefficient of association (19), $\phi(D_1, D_2) = 0$, and

Brown's coefficient of multilocus association (4), $Ass(D_1, D_2) = 0$. However, according to equation 1a, the value of correlation between these differentials is rather high, $Cor(D_1, D_2) = 0.81$.

Example 3.2. Let us consider a sample of isolates tested on two differentials, D_1 and D_2 , with virulence frequencies $p_1 = 0.05$ and $p_2 = 0.95$. Frequencies of four possible phenotypes are $q_{11} = p_1 \cdot p_2 = 0.0475$, $q_{10} = p_1(1 - p_2) = 0.0025$, $q_{01} = (1 - p_1)p_2 = 0.9025$, and $q_{00} = (1 - p_1)(1 - p_2) = 0.0475$, respectively, in the case of random distribution of virulences among isolates. There is no linkage between the differentials, $\phi(D_1, D_2) = 0$ and $Ass(D_1, D_2) = 0$. However, according to equation 1a, the correlation between these differentials is rather strong and negative, $Cor(D_1, D_2) = -0.81$.

Thus, examples 3.1 and 3.2 demonstrate that even in the case of the absolute equilibrium of genotypes with respect to two diallelic loci, the correlation between these loci may be positive or negative and rather strong.

Example 4. Let us consider a sample of isolates tested on two differentials, D_1 and D_2 , with the following frequencies of four possible phenotypes 11, 10, 01, and 00: $q_{11} = 0$, $q_{10} = 0.25$, $q_{01} = 0.25$, and $q_{00} = 0.5$, respectively. These two differentials are absolutely independent, $Cor(D_1, D_2) = 0$, whereas association among them equals $\phi(D_1, D_2) = -1/3$, and this value of linkage disequilibrium is significant for samples of more than 30 isolates.

Thus, the independence of differentiating characters does not necessarily imply linkage equilibrium within a population.

Example 5. Let us consider the following three samples of 1,000 isolates tested on a set of three differentials with virulence frequencies $p_1 = p_2 = p_3 = 0.1$ at each sample:

	1	2	3	Q_{ijk}		1	2	3	Q_{ijk}
	1	1	1	1		1	0	0	100
A	1	1	0	9	B	0	1	0	100
	1	0	1	9		0	0	1	100
	0	1	1	9		0	0	0	700
	1	0	0	81	C	1	2	3	Q_{ijk}
	0	1	0	81		1	1	1	100
	0	0	1	81		0	0	0	900
	0	0	0	729					

where Q_{ijk} is the number of individuals with ijk genotype pattern ($i, j, k = 0, 1$). Population A is in absolute equilibrium with no multilocus association, $Ass(A) = 0$, whereas linkage disequilibrium is obviously higher for population C compared with B , $Ass(C) = 2$ and $Ass(B) = -0.168$. Values of the measure of multilocus correlation for these populations equal $Cor(A) = 0.64$, $Cor(B) = 0.6$, and $Cor(C) = 1$. Therefore, increasing linkage disequilibrium does not necessarily imply a higher level of multilocus correlation (populations A and B versus populations A and C or B and C).

Thus, results of example 5 confirm once again that in general there is no relationship between multilocus correlation and multilocus association, and these two parameters reflect different aspects of the population's multilocus structure.

DIVERSITY AND MEASURE OF MULTILOCUS CORRELATION

Is there any relation between the measure of multilocus correlation and different measures of diversity within plant pathogen populations? We consider the following three commonly used diversity measures of different natures: (i) Shannon's index (15) in its normalized form (16), Sh , which is based on genotype (pathotype) frequencies; (ii) the Nei measure of gene diversity (13), H_S , which is based on frequencies of different alleles (virulence frequencies); and (iii) recently developed, Kosman's index (9), KW , which takes into account both genotype (pathotype) and allele (virulence) structure of populations.

Example 6. Let A_i be a population of plant pathogen which is presented by a sample of $2n$ isolates tested on a set of four

differentials and consists of i isolates with phenotype 1100 and $2n - i$ isolates with phenotype 0011, $0 \leq i \leq n$:

Population A_i				
1	1	1	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	1	1	0	0
$i + 1$	0	0	1	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$2n$	0	0	1	1

The measure of multilocus correlation for all these populations equals -1 , $Cor(A_i) = -1$, $0 \leq i \leq n$. It is easy to show that the Kosman, the Nei, and the Shannon normalized diversity indices give the following values:

$$KW(A_i) = \frac{i}{n}, \quad H_s(A_i) = \frac{i \cdot (2n-i)}{2n^2}, \quad \text{and} \quad Sh(A_i) = 1 - \frac{i \cdot \ln i + (2n-i) \cdot \ln(2n-i)}{2n \cdot \ln(2n)}$$

respectively, when $1 \leq i \leq n$, and $KW(A_0) = H_s(A_0) = Sh(A_0) = 0$ for population A_0 , which is limited to a single phenotype 0011. Then the following inequalities are satisfied:

$$\begin{aligned} 0 &= KW(A_0) < KW(A_1) < KW(A_2) < \dots < KW(A_n) = 1 \\ 0 &= H_s(A_0) < H_s(A_1) < H_s(A_2) < \dots < H_s(A_n) = 0.5 \\ 0 &= Sh(A_0) < Sh(A_1) < Sh(A_2) < \dots < Sh(A_n) = \ln 2 / \ln(2n) \end{aligned}$$

where 1 is the maximum of Kosman's index, 0.5 is the maximum of Nei's index in the case of diallelic loci, and $\ln 2 / \ln(2n)$ is the maximum value of Shannon's normalized index for samples of $2n$ isolates, which consist of two phenotypes. This means that the measures of diversity within populations increase from their absolute minimum 0 (for A_0) to their maximum (for A_n) despite the same minimal value of the measure of multilocus correlation for all considered populations.

Example 7. Let B_j be a population of plant pathogen which is presented by a sample of four isolates tested on a set of $2k$ differentials: two isolates are virulent on the first j differentials ($0 \leq j \leq k$) and avirulent on the rest $2k - j$ differentials; on the contrary, two other isolates are avirulent on the first j differentials and virulent on the others. Thus, the response patterns of population B_j ($0 \leq j \leq k$) are as follows:

Population B_j						
1	2	...	j	$j + 1$	...	$2k$
1	1	...	1	0	...	0
1	1	...	1	0	...	0
0	0	...	0	1	...	1
0	0	...	0	1	...	1

The Kosman diversity index for B_j is absolutely maximal, $KW(B_j) = 1$, $0 \leq j \leq k$. The Shannon normalized index equals its maximum for populations of four isolates, which consist of two phenotypes $Sh(B_j) = 0.5$, $0 \leq j \leq k$. Nei's diversity index also has its maximal value in the case of diallelic loci, $H_s(B_j) = 0.5$. The measure of multilocus correlation for population B_j equals $Cor(B_j) = 1 - 2 \cdot j/k$, $0 \leq j \leq k$. Then, the following inequalities are met by $1 = Cor(B_0) > Cor(B_1) > Cor(B_2) > \dots > Cor(B_k) = -1$.

Thus, the measure of multilocus correlation may range from -1 to 1 for populations with the same maximal value of diversity within them.

Example 8. Let us consider seven populations ($P_1, P_2, P_3, P_4, P_5, P_6$, and P_7) each of which consists of four isolates tested on six differentials with phenotypes as follows:

P_1	P_2	P_3	P_4	P_5	P_6	P_7
100000	100000	100000	110000	110000	100000	100000
011111	100000	111111	011111	110000	110000	110000
011111	011111	011111	011111	001111	011111	011111
011111	011111	011111	011111	001111	011111	001111

The calculated values of the Kosman, the Shannon normalized, and the Nei diversity indices and the measures of multilocus correlation and association for these populations are presented in the following table:

	P_1	P_2	P_3	P_4	P_5	P_6	P_7
KW	1/2	1	7/12	5/12	1	11/12	1
Sh	$1 - \frac{3 \ln 3}{4 \ln 4}$	1/2	3/4	$1 - \frac{3 \ln 3}{4 \ln 4}$	1/2	3/4	1
H_s	3/8	1/2	19/48	15/48	1/2	23/48	1/2
Cor	1/3	1/3	1/2	1/6	-1/3	1/6	0
Ass	5	5	3.736	4	5	3.790	3.333

Relations between these diversity indices and the measure of overall correlation between differentials are expressed by the following inequalities: $KW(P_1) < KW(P_3) < KW(P_2)$ and $H_s(P_1) < H_s(P_3) < H_s(P_2)$, while $Cor(P_1) < Cor(P_3)$ but $Cor(P_3) > Cor(P_2)$; and $Cor(P_4) < Cor(P_2) < Cor(P_3)$, while $KW(P_4) < KW(P_2)$ but $KW(P_2) > KW(P_3)$, and $H_s(P_4) < H_s(P_2)$ but $H_s(P_2) > H_s(P_3)$.

These relations mean that the increase of Kosman's and Nei's diversity indices do not necessarily imply the increase of the measure of multilocus correlation and vice versa.

The Shannon diversity index and the measure of multilocus correlation for populations P_5, P_6 , and P_7 satisfy the following inequalities: $Sh(P_5) < Sh(P_6) < Sh(P_7)$, while $Cor(P_5) < Cor(P_6)$ but $Cor(P_6) > Cor(P_7)$; and $Cor(P_5) < Cor(P_7) < Cor(P_6)$, while $Sh(P_5) < Sh(P_7)$ but $Sh(P_7) > Sh(P_6)$.

Therefore, the increase of Shannon's diversity index does not necessarily imply the increase of the measure of overall correlation between differentials and vice versa.

In addition, comparison between the diversity indices and the measure of multilocus association does not reveal any dependence between them. For example, $KW(P_4) < KW(P_1)$ and $Ass(P_4) < Ass(P_1)$, whereas $KW(P_1) < KW(P_6)$ but $Ass(P_1) > Ass(P_6)$.

Thus, there are not any relations between different indices of diversity within population and the measures of multilocus correlation and association. All these parameters of a pathogen population are independent.

APPLICATIONS OF THE MEASURE OF MULTILOCUS CORRELATION

The measure of multilocus correlation is a parameter that describes relations between diallelic loci with respect to the given population. Therefore, this measure would naturally be used for comparisons between different sets of differentials with respect to the same population. In particular, independence or correlation between two differentiating characters can be tested by this measure (equation 1). If the value of correlation between two given differentiating factors is relatively close to 1 or to -1 , then a positive or a negative correlation, respectively, is observed; otherwise, these two differentials are considered independent (or more correctly, not significantly correlated).

Clustering. The bootstrap statistical method (5,6) is a widely used procedure for clustering isolates of a plant pathogen tested on a set of differentiating factors. The independence of differentiating characters is a necessary condition for the bootstrapping across the characters to be valid. Therefore, the measure of overall correlation between differentials can help in a search for the optimal subset, which consists of independent pairs of differentials. Then a bootstrap involving sampling differentials from this optimal subset at random would have been more appropriate than re-sampling from the entire original set of differentials. The inferences from bootstrap analyses will be distorted if the differential factors used to characterize the isolates are significantly correlated. This argument is especially important for plant pathogen populations that reproduce asexually. The natural dependence of markers in the case of asexual reproduction make application of

the bootstrap method across differentiating characters much more problematic than in other cases.

Example 9. Let us consider the following sample of 2,000 isolates tested on a set of three differentials with virulence frequencies $p_1 = p_2 = 0.05$ and $p_3 = 0.4$:

	1	2	3	Q_{ijk}
P	1	1	1	2
	1	1	0	3
	1	0	1	38
	0	1	1	38
	1	0	0	57
	0	1	0	57
	0	0	1	722
	0	0	0	1,083

where Q_{ijk} is the number of individuals with ijk genotype pattern ($i, j, k = 0, 1$). Population P is in absolute equilibrium with respect to all pairs and the entire set of differentials, $Ass(P) = 0$. The value of multilocus correlation, $Cor(P) = 0.39$, is far from 0, which might result from rather strong positive correlation between some pairs of differentials. Testing of three possible pairs of differentials shows a high level of correlation between the first two differentials, $Cor(D_1, D_2) = 0.81$, whereas two other pairs are close to independence, $Cor(D_1, D_3) = Cor(D_2, D_3) = 0.18$.

In the case of cluster analysis of these isolates, statistical support of the obtained groups can be realized by bootstrapping across the characters. Each bootstrap sample consists of a new data table with the same set of isolates, but with some of the original characters duplicated a few times and others dropped by the process of random sampling characters with replacement. Then, only about 33% of fictional data tables would include the third differential, D_3 , and therefore, the cluster, which includes 722 isolates with the pattern 001, has weak statistical support (33 appearance from 100). However, this result is not valid because the patterns of the first two differentials, D_1 and D_2 , are rather dependent, $Cor(D_1, D_2) = 0.81$, despite the absolute equilibrium of the given individuals with respect to these two differentials, $Ass(D_1, D_2) = 0$ and $\phi(D_1, D_2) = 0$. The more suitable approach is to consider any independent subset of differentials, either $\{D_1, D_3\}$ or $\{D_2, D_3\}$, and then to apply the bootstrap method. Then statistical support of the above-mentioned cluster would be much stronger (50%).

Thus, the measure of overall correlation can help to find the maximal subset of independent (not significantly correlated) differentials even in the case of the absolute multilocus equilibrium.

Using the bootstrap method across the characters for confident clustering solutions by dendrograms is not well-founded and not an accurate procedure in its practical realization. The general discussion of this problem is beyond the objectives of this paper. The only point to touch on here is how to determine if two characters are significantly correlated. In other words, the question is how to set the significance level of correlation. It would be reasonable to tie this with the minimal size of clusters we want to detect. For example, if the minimal size of clusters is 10% of the original sample size, then the significance level of correlation has to be 0.2. This means that two characters will be considered strongly correlated if the measure of correlation between them is greater than 0.8 (at least 90% of isolates have the same virulence type on both differentials D_1 and D_2 as $Cor(D_1, D_2) > 0.8$, and due to equation 1, $d(D_1, D_2)/n < 0.1$) or less than -0.8 (at least 90% of isolates do not share the same virulence type on both differentials), and then one of these characters should be discarded. Otherwise, if the measure of correlation ranges between -0.8 and 0.8 , the characters are considered independent and they may be involved in the optimal set of differentials for the bootstrapping. In the case of example 9, one of the differentials D_1 or D_2 has to be removed, and the remaining two differentials may be involved in the bootstrapping. If the significance level of correlation equals 0.1, then all three differentials may be included in the optimal set

of characters because the measure of correlation scores values between -0.9 and 0.9 for all pairs of the differentials, and they are considered independent. The maximum theoretical value of the significance level is 1.

With research data, it is important to determine the significance of different measurements. This point is more philosophical and methodological than a purely computational (statistical) problem. In general, formal criterion cannot be applied for deciding when differentials are independent. Choice of a suitable significance level is subject to the character and goals of a specific study and may depend on the specific host–pathogen system.

Diversity within population. The measure of multilocus correlation also can be used as an additional parameter for the study of diversity within pathogen populations. The Shannon, Nei, and Kosman indices are usually used for measuring diversity within plant pathogen populations. The Shannon index may be chosen when comparing species diversity within ecological communities, but for populations within a species, it is better to take genetic similarities into account. If just the phenotype (genotype) structure of a pathogen population is studied, then the Shannon index can be applied under the assumption that nonidentical phenotypes are equally distinct. The Nei index is based on virulence (band) frequencies regardless of phenotypes (genotypes). Nei’s diversity is good for populations of randomly mating organisms, but for most plant-pathogenic fungi that have prominent asexual stages in their life cycles, the Kosman index is better because it accounts for both phenotype (genotype) structure of the population and genetic similarities among clonal lineages within the populations. However, even populations that have identical Kosman index values may have subtle differences in genetic diversity that can be revealed by the multilocus correlation parameter. Thus, the multilocus correlation parameter ordinarily would be used only after calculation of the Nei index or the Kosman index for comparing diversity among populations.

Example 10. Let us consider two populations, each of which consists of four isolates tested on six differentials with phenotypes as follows:

A	B
1 0 0 0 0	1 1 1 0 0 0
1 0 0 0 0	1 1 1 0 0 0
1 1 1 1 1	1 0 0 1 1 1
0 1 1 1 1	0 0 0 1 1 1

The Kosman, the Shannon normalized, and the Nei diversity indices are equal for these populations: $KW(A) = KW(B) = 11/12$, $Sh(A) = Sh(B) = 3/4$, and $H_S(A) = H_S(B) = 23/48$. The Simpson diversity index (17) also is equal for these populations (5/8) because they both consist of three pathotypes with frequencies of 0.5, 0.25, and 0.25. The virulence frequencies on the corresponding differentials also are equal for populations A and B , which implies the same value for another Kosman’s genetic diversity H_{KDiv} (11) and Muller’s index (11,12) for these populations. Therefore, according to all mentioned parameters, populations A and B are considered equally diverse. However, the measure of multilocus correlation scores different values for these populations: $Cor(A) = 1/2$ and $Cor(B) = -5/6$, and $Cor(B) < Cor(A)$. This means that in the case of population A , the differential characters have more uniform patterns across the entire sample of isolates (columns of data matrix). Therefore, it is natural to assume that population A is less diverse than population B .

Generally, let us consider a dichotomous pattern of any individual, i , tested on a set of k differentiating factors with positive responses (1s) on m of them, for instance

	1	2	...	m	$m + 1$	...	k
i	1	1	...	1	0	...	0

The relative complexity, C , of an individual is determined as a number of positive responses (1s) on a set of differentials divided

by the total number of differentiating factors, $C(i) = m/k$. The relative complexity varies between 0 and 1.

We define the following measure of uniformity for any individual, i , as

$$U(i) = 1 - 2 \cdot \min\{C(i), 1 - C(i)\} \quad (3)$$

This measure ranges from 0 to 1. If there are only 1s (or 0s) in the response pattern, then the uniformity is maximal and equals 1. The uniformity reaches its minimum value of 0 if the numbers of positive (1s) and negative (0s) responses are equal. For example, for the following three isolates (i_1 110000, i_2 111000, and i_3 111110), the relative complexity equals $C(i_1) = 2/6 = 1/3$, $C(i_2) = 1/2$, and $C(i_3) = 5/6$, respectively, and the corresponding values of the uniformity measure are

$$U(i_1) = 1 - 2 \min\left\{\frac{1}{3}, \frac{2}{3}\right\} = \frac{1}{3}, U(i_2) = 1 - 2 \min\left\{\frac{1}{2}, \frac{1}{2}\right\} = 0, \text{ and } U(i_3) = 1 - 2 \min\left\{\frac{5}{6}, \frac{1}{6}\right\} = \frac{2}{3}$$

Therefore, the third and the second isolates are the most and the least uniform, respectively, among these three isolates.

If any sample from population P consists of n individuals i_j ($j = 1, 2, \dots, n$), then we define the measure of uniformity for P as the average of the uniformity values (equation 3) across individuals

$$U(P) = \frac{1}{n} \cdot \sum_{j=1}^n U(i_j) \quad (4)$$

In the case of example 10, the uniformity of population A , $U(A) = 3/4$, is greater than that of B , $U(B) = 1/12$. This fact explains why population A can be considered as less diverse than population B .

Example 11. Let us consider two populations, each of which consist of four isolates tested on six differentials with phenotypes as follows:

A	B
1 1 0 0 0 0	1 1 1 0 0 0
1 1 0 0 0 0	1 1 1 0 0 0
0 0 1 1 0 0	0 0 0 1 0 0
0 0 1 1 0 0	0 0 0 1 0 0

The Kosman, the Shannon normalized, and the Nei diversity indices are equal for these populations: $KW(A) = KW(B) = 2/3$, $Sh(A) = Sh(B) = 1/2$, and $H_3(A) = H_3(B) = 1/3$. The Simpson diversity index (17) also is equal for these populations (1/2) because they both consist of two pathotypes with frequencies of 0.5. The virulence frequencies on the corresponding differentials also are equal for populations A and B , which implies the same value of another Kosman's genetic diversity H_{KDiv} (11) and Muller's index (11,12) for these populations. Therefore, according to all mentioned diversity indices, populations A and B are considered equally diverse. The measure of multilocus correlation also scores equal values for these populations, $Cor(A) = Cor(B) = -1/3$, and they are equally uniform as well, $U(A) = U(B) = 1/3$. Thus, none of the considered parameters makes a difference between diversities of populations A and B . Meanwhile, population B can be considered more diverse than A due to more differences between the uniformity of individuals and the average uniformity of B . On the other hand, population A can be considered more diverse than B because differentials' patterns are identical for A and B , but distribution of these patterns is more even in the case of population A .

We define the absolute deviation of an individual's uniformity from the average uniformity for an arbitrary population P by the formula

$$dU(P) = \frac{1}{n} \cdot \sum_{j=1}^n |U(i_j) - U(P)|$$

This parameter measures variance of individuals' uniformity for a given population. It is natural to assume that a population with a higher level of this absolute deviation is more variable in the case

of the same value of the population uniformity. The value of absolute deviation of uniformity equals $dU(A) = 0$ and $dU(B) = 1/3$ for populations A and B , respectively, from example 11. Therefore, according to this parameter, population B is more diverse than A .

Another parameter that can be introduced for comparison between populations is the measure of average dissimilarity between patterns of differentials:

$$Dis(P) = \frac{1}{k^2} \cdot \sum_{i,j=1}^k d(D_i, D_j)$$

where D_i ($i = 1, 2, \dots, k$) are k differentiating factors for individuals from population P and $d(D_i, D_j)$ is the distance between patterns of differentials i and j which equals the number of individuals with different responses to D_i and D_j . It is natural to assume that a population with a higher level of the average dissimilarity between differentials is more diverse. Therefore, population A is more diverse than B (example 11) according to this parameter because $Dis(A) = 16/9$ is greater than $Dis(B) = 14/9$. The effect of the average dissimilarity between differentials on diversity within populations A and B compensates for the effect of the variance of individuals' uniformity for these populations, which can explain why the measure of overall correlation between differentials does not help to distinguish between diversities within populations A and B .

Thus, the measure of multilocus correlation can be used for measuring diversity within a population if all other diversity indices fail to draw a distinction between diversities within different populations. In such a case, the higher the overall correlation, the less diverse the population.

CONCLUSION

The measure of multilocus correlation is a new parameter for the study of multilocus structure of populations, which describes different aspects of this structure compared with the commonly used measure of multilocus association and ϕ coefficient of association.

The measure of multilocus correlation can be used for analysis of independence of differentiating characters for a given population—the necessary condition of valid applications of the bootstrap method across differentials.

If the standard diversity indices do not distinguish between populations, then the measure of multilocus correlation is suitable for disclosing slight differences in diversity within populations.

Thus, the measure of overall correlation can be a useful and effective tool for study and comparison between plant pathogen populations.

ACKNOWLEDGMENTS

I thank K. J. Leonard (University of Minnesota) for critical reading of the manuscript and numerous helpful comments and suggestions.

LITERATURE CITED

1. Bellman, R., Cooke, K. L., and Lockett, J. A. 1970. Algorithms, Graphs and Computers. Academic Press, New York.
2. Borchardt, D. S., Welz, H. G., and Geiger, H. H. 1998. Genetic structure of *Setosphaeria turcica* populations in tropical and temperate climates. Phytopathology 88:322-329.
3. Borie, B., Jacquot, L., Jamaux-Despréaux, I., Larignon, P., and Péros, J.-P. 2002. Genetic diversity in populations of the fungi *Phaeoemoniella chlamydospora* and *Phaeoacremonium aleophilum* on grapevine in France. Plant Pathol. 51:85-96.
4. Brown, A. H. D., Feldman, M. W., and Nevo, E. 1980. Multilocus structure of natural populations of *Hordeum spontaneum*. Genetics 96:523-536.
5. Diaconis, P., and Efron, B. 1983. Computer-intensive methods in statistics. Sci. Am. 249:116-130.

6. Felsenstein, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. *Evolution* 39:783-791.
7. Gale, L. R., Chen, L.-F., Hernick, C. A., Takamura, K., and Kistler, H. C. 2002. Population analysis of *Fusarium graminearum* from wheat fields in eastern China. *Phytopathology* 92:1315-1322.
8. Kolmer, J. A., and Liu, J. Q. 2000. Virulence and molecular polymorphism in international collections of the wheat leaf rust fungus *Puccinia triticina*. *Phytopathology* 90:427-436.
9. Kosman, E. 1996. Difference and diversity of plant pathogen populations: A new approach for measuring. *Phytopathology* 86:1152-1155.
10. Liu, J. Q., and Kolmer, J. A. 1998. Molecular and virulence diversity and linkage disequilibria in asexual and sexual populations of the wheat leaf rust fungus, *Puccinia recondita*. *Genome* 41:832-840.
11. Manisterski, J., Eyal, Z., Ben-Yehuda, P., and Kosman, E. 2000. Comparative analysis of indices in the study of virulence diversity between and within populations of *Puccinia recondita* f. sp. *tritici* in Israel. *Phytopathology* 90:601-607.
12. Müller, K., McDermott, J. M., Wolfe, M. S., and Limpert, E. 1996. Analysis of diversity in populations of plant pathogens: The barley powdery mildew pathogen across Europe. *Eur. J. Plant Pathol.* 102:385-395.
13. Nei, M. 1973. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. USA* 70:3321-3323.
14. Salamati, S., Zhan, J., Burdon, J. J., and McDonald, B. A. 2000. The genetic structure of field populations of *Rhynchosporium secalis* from three continents suggests moderate gene flow and regular recombination. *Phytopathology* 90:901-908.
15. Shannon, C. E., and Weaver, W. 1949. *The Mathematical Theory of Communication*. University of Illinois Press, Urbana.
16. Sheldon, A. L. 1969. Equitability indices: Dependence on the species count. *Ecology* 50:466-467.
17. Simpson, E. H. 1949. Measurement of diversity. *Nature* 163:688.
18. Slatkin, M. 1994. Linkage disequilibrium in growing and stable populations. *Genetics* 137:331-336.
19. Sokal, R. R., and Rohlf, F. J. 1995. *Biometry*. W. H. Freeman, New York.
20. Weir, B. S. 1979. Inferences about linkage disequilibrium. *Biometrics* 35:235-254.
21. Welz, H. G., and Leonard, K. J. 1995. Gametic phase disequilibria in populations of race 2 and race 3 of *Cochliobolus carbonum*. *Eur. J. Plant Pathol.* 101:301-310.
22. Zhan, J., Mundt, C. C., and McDonald, B. A. 1998. Measuring immigration and sexual reproduction in field populations of *Mycosphaerella graminicola*. *Phytopathology* 88:1330-1337.