Molecular Ecology (2005) **14**, 415–424 doi: 10.1111/j.1365-294X.2005.02416.x

Similarity coefficients for molecular markers in studies of genetic relationships between individuals for haploid, diploid, and polyploid species

E. KOSMAN* and K. J. LEONARD†

**Institute for Cereal Crops Improvement (ICCI), The George S. Wise Faculty for Life Sciences Tel Aviv University Tel Aviv 69978, Israel, †Department of Plant Pathology, University of Minnesota, St. Paul, MN 55108, USA*

Abstract

Determining true genetic dissimilarity between individuals is an important and decisive point for clustering and analysing diversity within and among populations, because different dissimilarity indices may yield conflicting outcomes. We show that there are no acceptable universal approaches to assessing the dissimilarity between individuals with molecular markers. Different measures are relevant to dominant and codominant DNA markers depending on the ploidy of organisms. The Dice coefficient is the suitable measure for haploids with codominant markers and it can be applied directly to (0,1)-vectors representing banding profiles of individuals. None of the common measures, Dice, Jaccard, simple mismatch coefficient (or the squared Euclidean distance), is appropriate for diploids with codominant markers. By transforming multiallelic banding patterns at each locus into the corresponding homozygous or heterozygous states, a new measure of dissimilarity within locus was developed and expanded to assess dissimilarity between multilocus states of two individuals by averaging across all codominant loci tested. There is no rigorous well-founded solution in the case of dominant markers. The simple mismatch coefficient is the most suitable measure of dissimilarity between banding patterns of closely related haploid forms. For distantly related haploid individuals, the Jaccard dissimilarity is recommended. In general, no suitable method for measuring genetic dissimilarity between diploids with dominant markers can be proposed. Banding patterns of diploids with dominant markers and polyploids with codominant markers represent individuals' phenotypes rather than genotypes. All dissimilarity measures proposed and developed herein are metrics.

Keywords: assignment problem, codominant markers, diversity, dominant markers, population genetics

Received 26 August 2004; revision received 20 October 2004; accepted 20 October 2004

Introduction

Two approaches are commonly used in studies of genetic diversity within and among populations or groups of individuals. In the first, allele frequencies over a number of polymorphic loci are determined, and parameters based on the allele frequencies are used for partitioning genetic variation into components for variation within and between units. This approach may be chosen when dominant markers such as RAPDs (rapid analyses of polymorphic DNA), AFLPs (amplified fragment length polymorphisms) and ISSRs (inter simple sequence repeats) are applied to haploid individuals or codominant markers such as allozymes, RFLPs (restriction fragment length polymorphism) and SSRs (simple sequence repeats) (microsatellites) are used with haploid or diploid species with the assumption of no linkage between loci. With dominant markers, individuals that are heterozygous for a DNA band at a specific position cannot be distinguished with certainty from individuals that are heterozygous for that band. With sexually reproducing organisms in randomly mating populations, the allele frequencies can be inferred by assuming Hardy–Weinberg equilibrium (HWE) and independent assortment of genes. The

Correspondence: Dr E. Kosman, Fax: 972-3-6407857, E-mail: kosman@post.tau.ac.il

second approach, which is applied with all types of markers and organisms, is based on comparisons of individual genotypes within and between populations. A genetic dissimilarity matrix constructed from all possible pairwise combinations of individuals is used for characterizing population structure based on relative affinities of each individual to all other individuals tested. This approach requires proper methods for assessing dissimilarity between individuals, and it is particularly useful in the case of possible linkages between different loci. The choice of a suitable index of similarity is a very important and decisive point for determining true genetic dissimilarity between individuals, clustering, analysing diversity within populations and studying relationship between populations, because different dissimilarity indices may yield contrary outcomes. Many researchers have preferred for various well-documented reasons to use the second approach either alone or in combination with the first approach. However, the bases for choosing the most appropriate coefficient of dissimilarity depending on type of marker and ploidy of the organism in question, which is the subject of this study, have not received sufficient attention in published research articles.

Molecular markers are commonly used to characterize genetic diversity within or between populations or groups of individuals because they typically detect high levels of polymorphism. Furthermore, RAPDs and AFLPs are efficient in allowing multiple loci to be analysed for each individual in a single gel run. In analysing banding patterns of molecular markers, the data typically are coded as (0,1)-vectors, 1 indicating the presence and 0 indicating the absence of a band at a specific position in the gel. With diploid organisms and codominant markers, such as allozymes, RFLPs or SSRs, the banding patterns may be translated to homozygous or heterozygous genotypes at each locus, and the allelic structure derived is utilized for comparison between individuals (Peakall *et al*. 1995; Smouse & Peakall 1999; Maguire *et al*. 2002). More often, however, the binary patterns obtained are used directly in comparisons of similarity of individuals.

We study the general case of research on structure and genetic diversity in populations for which pedigree information is absent. We will not consider special cases, where the ancestries of progeny have already been traced for all individuals, and the theory of kinship coefficients might be applied to quantify genetic relatedness between pairs of individuals. Also, we do not intend that all of our analyses be extended to comparisons among distantly related species for which the use of dominant molecular markers may be questionable.

Several measures including the Dice (Nei and Li), Jaccard, and simple match (or the squared Euclidean distance) coefficients are commonly employed in the analyses of similarity of individuals (binary patterns) in the absence of knowledge of ancestry of all individuals in the populations. These similarity coefficients are defined differently and therefore they may yield different results for both the qualitative and quantitative relationships between individuals. Although these coefficients may not yield identical results, most published studies do not offer any rationale to support the choice of the coefficient that was used in relation to the type of marker evaluated or the ploidy and mating system of the organism being studied. For example, the May 2004 issue of *Molecular Ecology* contains three papers (Brouat *et al*. 2004; Mock *et al*. 2004; Pannebakker *et al*. 2004), where all three dissimilarity measures discussed (the Dice, Jaccard and simple mismatch coefficients or their equivalents) were used for analyses of the same type of data (AFLP profiles) without any rationale being offered by the authors for their choice of the measure they used. Moreover, Brouat *et al*. (2004) inconsistently employed two qualitatively different measures with the same data: the squared Euclidean distance for AMOVA and the Dice coefficient for constructing neighbour-joining tree. On the other hand, the same coefficients have often been used for both dominant (RAPD, AFLP and ISSR) and codominant (allozymes, RFLP and SSR) markers and without regard for whether the species being studied are haploid, diploid or polyploid, or the degree of genetic recombination or heterozygosity expected from its mating system. Each of these factors may influence how accurately the direct application of a given similarity coefficient to the (1,0)-vectors will reflect the true genetic similarity of any pair of individuals. In most published studies, the similarity coefficient used was apparently chosen simply because it was used in an earlier publication or it is available in the software package used to analyse the data. In some cases, two or three similarity coefficients are used with the same data set (Cordeiro *et al*. 2003: Jaccard and Dice coefficients for SSRs with polyploids; Kumar *et al*. 1999: Dice and simple match coefficients for DNA fingerprints with haploids; Pei & Ruiz 2000: Jaccard, Dice and simple match coefficients for AFLPs with asexual dikaryons that function as diploids with one pair of haploid nuclei per cell) with the expectation that if the results are robust, the different coefficients should reveal essentially the same patterns of diversity. If two similarity coefficients reveal somewhat different patterns of relationships between individuals, there is rarely any rationale presented to suggest which pattern is more valid, and often only one of the patterns is presented in the publication.

Another problem appears to stem from the absence of absolute separation between the two approaches for diversity analysis. For example, it was proved (Kosman 2003) that the Nei heterozygosity H_E and the mean pairwise dissimilarity between individuals with respect to the simple mismatch coefficient are identical measures of diversity within populations. Therefore, applications of the Dice or

Jaccard dissimilarities together with the Nei diversity H_E lead to the hidden inconsistency of analysis.

As a general rule, we should expect an appropriate similarity coefficient to produce a consistent measure of the proportion of differentiating factors showing similarity between any pair of individuals relative to the total number of factors in which differences could have been detected if the individuals showed no detectable similarity. That is, the similarity coefficient employed should accurately reflect our best understanding of the phenotypes observed and the genetic basis for them. The main objectives of this study are to (i) provide a rationale for choosing the method that best reflects the true phenotypic and genetic similarity between pairs of individuals; and (ii) develop new methods and select the most appropriate indices of similarity for comparison between individuals for the following four systems: codominant markers — haploid organisms, codominant markers — diploid organisms, dominant markers — haploid organisms, and dominant markers — diploid organisms. To meet these objectives in studies of population structure within species, it is often not possible to define lines of descent of individuals from common ancestors as might be performed in evolutionary analyses of phylogenetically related species. Thus, we consider that the presence of a specific allele in different populations could be the result of migration between populations and not necessarily proof of shared ancestry among the founders of the different populations. Our approach does not assume that each population arose from a single progenitor or that different populations must have remained totally isolated from one another from the beginning of their inception.

Definitions and comparison of commonly used measures of similarity

Any two individuals i_1 and i_2 tested for molecular markers may be represented by their binary patterns derived from the corresponding banding patterns of their DNA markers. These binary patterns are (0,1)-vectors, for which 1 designates presence and 0 designates absence of a band at some position. There are two obvious options for determining size of these (0,1)-vectors. If only two individuals are compared, the vector size might equal the number of positions at which a band appears for at least one of the two. However, if several pairs of individuals from any sample are compared, the size of the corresponding (0,1)-vectors should equal the total number of band positions (or possible band positions) where a band appears in at least one individual of the sample. That is, the size of the (0,1)-vectors should reflect the total number of polymorphic loci represented in the sample.

We denote a = number of positions with shared bands (1s) for both individuals; b = number of positions where individual i_1 has a band, but i_2 does not; c = number of positions where individual i_2 has a band, but i_1 does not; and n = the size of (0,1)-vectors. The following measures of similarity (Sneath & Sokal 1973) are usually used for comparison between individuals: Jaccard's coefficient $J(i_1, i_2) = a/(a + b + c)$ (Teulat *et al.* 2000 for AFLPs; Anthony *et al.* 2002 for AFLPs and SSRs; Cordeiro *et al.* 2003 for SSRs; Schönswetter *et al.* 2003 AFLPs), Dice's coefficient $D(i_1, i_2) = 2a/(2a + b + c)$ (Barth *et al.* 2002 for ISSR; Belaj *et al.* 2003 for RAPDs, AFLPs and SSRs; Cordeiro *et al.* 2003 for SSRs; Dearborn *et al.* 2003 for AFLPs), and the simple match coefficient $M(i_1, i_2) = (n - b - c)/n$ (Peakall *et al.* 1995 for RAPDs; Teulat *et al.* 2000 for SSRs). Also, $j(i_1, i_2) = 1 - J(i_1, i_2) = (b + c)/(a + b + c)$, $d(i_1, i_2) = 1 - D(i_1, i_2) = (b + c)/(2a + b + c)$ and $m(i_1, i_2) = 1 - M(i_1, i_2) = (b + c)/n$ are the Jaccard, Dice and simple mismatch coefficients of dissimilarity between individuals i_1 and i_2, respectively. Values for all these coefficients range between 0 and 1. Sometimes, the Euclidean distance $e(i_1, i_2)$ between isolates or its squared value $e^2(i_1, i_2) = b + c$ are used (Huff *et al.* 1993 for RAPDs; Torimaru *et al.* 2003 for RAPDs; Bleeker 2003 for AFLPs). The Dice coefficient of similarity is frequently referred to as the measure of genetic similarity of Nei & Li (1979). For a given data set, the corresponding values of Jaccard's dissimilarity are always greater than those of the Dice dissimilarity and the simple mismatch coefficient. On the other hand, values of the Dice dissimilarity may be greater or smaller than the corresponding values of the simple mismatch coefficient depending on whether the number of positions with shared bands a is less or greater than the number of positions with shared absence of bands $n - (a + b + c)$, respectively.

Jaccard's and Dice's measures do not account for $n - (a + b + c)$ factors where both individuals i_1 and i_2 respond negatively, and they are bound with the following formulae: $D = 2J/(1 + J)$ or $d = j/(2 - j)$. The simple mismatch coefficient and the normalized squared Euclidean distance are identical measures because of the equation $m = e^2/n$. One can prove that Jaccard's and Dice's coefficients reflect relation of proximity between two arbitrary pairs of individuals in a qualitatively similar way, as do the simple mismatch coefficient and the Euclidean distance. However, relations of proximity between individuals measured by the simple mismatch coefficient and the Euclidean distance generally differ from those measured by Jaccard's and Dice's indices. We will not further consider the Euclidean distance with nominal data.

There are significant qualitative differences in definitions of the simple match coefficients vs. the Jaccard and Dice similarities that may impose some restrictions on application of these indices to specific data. The simple match coefficient takes into account both the shared 0s (absence of a band) and shared 1s (presence of a band) as factors that contribute to similarity between two individuals. The

Jaccard coefficient of similarity considers only shared 1s as contributing to the similarity of individuals and disregards shared 0s. The Dice coefficient of similarity also ignores shared 0s, and in addition, it differs quantitatively from Jaccard's similarity in that the Dice measure of similarity attaches more importance to the factors with positive response for both individuals (shared 1s) than to those with positive response in only one individual or the other, which is expressed by the product $2a$ compared to b and c in the definitions.

According to the conservative approach, DNA fingerprint similarity is generally defined as the fraction of observed bands that are shared by two individuals. Even if such approach is valid and shared 0s are ignored, the question is how to make a suitable choice between the Jaccard or Dice coefficients. Furthermore, it is possible to create an infinite series of potentially appropriate indices, which are generalizations of the Dice coefficient.

Because the Dice coefficient attaches more importance to the shared presence of 1s (bands) then we can attribute an increasing weight to positive responses shared by both individuals if we introduce a series of similarity indices $K_s(i_1, i_2) = sa/(sa + b + c)$ for $s = 1, 2, 3, \dots$, where $K_1 = J$ (Jaccard's similarity) and $K_2 = D$ (Dice's similarity). It is easy to see that $K_s < K_t$ for all $s < t$ and the corresponding dissimilarities $k_s = 1 - K_s$ and $k_t = 1 - K_t$ satisfy the inequality $k_s > k_t$. In particular, $J < D$ and $j > d$.

Nei & Li (1979) proposed their index for measuring similarity between two populations as the ratio of two expectations — the number of shared bands for a pair of randomly drawn individuals (one from each population), and the number of bands exhibited by a randomly sampled individual from the pooled population. This index being applied to two individuals = 'populations' might be considered as another interpretation of the Dice similarity if the expectations are the corresponding average values. The similarity of two individuals i_1 and i_2 is defined as the number of common bands in their fingerprint profiles, a, divided by the average number of fragments exhibited by both individuals,

$$\hat{n} = \frac{(a + b) + (a + c)}{2} = \frac{2a + b + c}{2},$$

where $n_1 = a + b$ and $n_2 = a + c$ are the numbers of bands obtained for individuals i_1 and i_2, respectively. The average number of bands $\hat{n}$ is interpreted as the number of fragments for the common ancestor of the two individuals. If three (or more) individuals originated from the common ancestor, then according to the previous logic, the average number of fragments that appear for these individuals should estimate the number of bands expected for this ancestor. The problem is how to measure similarity between any two individuals from such a group. Should the number of shared bands be divided by the average number of bands for the three individuals compared? Probably not, because this measure might theoretically be greater than 1. Furthermore, Nei & Li (1979) based their approach on the assumption that differences in banding patterns arise as mutations from a common ancestor, whereas differences within populations may also reflect other factors such as founder effects and gene flow between populations. Thus, it is not so obvious that the widely used and recommended Dice (Nei and Li) coefficient of similarity is always a biologically suitable and correctly interpreted measure of similarity for comparing fingerprint profiles.

Codominant markers and similarity measures

With codominant markers, such as allozymes, RFLP and SSR, each recognizable allele at a given locus is ordinarily associated with a single band at a unique position in the gel. Thus, in the case of diploid organisms for a given locus a homozygote will have one band and a heterozygote will have two. Null alleles (no band) are rarely seen. Therefore, the shared absence of a band at a specific position should not be considered in measures of similarity with codominant markers. Clearly with codominant markers, the genetic similarities between pairs of individuals cannot be characterized simply in terms of the proportion of bands that are shared between two individuals. Also, if there are multiple alleles per locus, as expected for SSRs, which are highly variable, the total number of bands expressed by all the individuals in a sample will likely be much greater than the number of loci involved. Therefore, the banding profiles should be adjusted to represent the allelic patterns of individuals across all loci studied, and to represent the total number of loci and the number of shared alleles rather than the total number of bands and the number of shared bands, respectively, and the adjusted values should be employed for measuring similarity between individuals.

Haploid organisms

Let us consider a haploid individual i subject to genetic analysis by codominant molecular markers in n loci. We denote $r = r_j$ the number of alleles in a multiallelic locus f_j ($j = 1, 2, \dots, n$), and $S_j = \{s_{j1}, s_{j2}, \dots, s_{jr}\}$ is the set of these alleles. An individual i can be represented by its vector of states $i = (s_1, s_2, \dots, s_n)$, where s_j is one of the alleles (S_j-states) of locus f_j, $j = 1, 2, \dots, n$. If two individuals i_1 and i_2 are represented by their state-patterns $i_k = (s_{1k}, s_{2k}, \dots, s_{nk})$ for $k = 1, 2$, then it is natural to determine the similarity between these two individuals as the number of shared alleles across all loci considered, and to normalize this value by the total number of loci n. This measure of similarity between individuals is the simple match coefficient $M(i_1, i_2)$ in the case of multistate characters. Formally, for two alleles s_{ju} and s_{jv} of locus f_j we define $\delta(s_{ju}, s_{jv}) = 0$ if $u = v$, and $\delta(s_{ju}, s_{jv}) = 1$ if $u \neq v$ ($u, v = 1, 2, \dots, r$). Then the simple match coefficient $M\ (i_1, i_2)$ has the following form:

$$M(i_1, i_2) = \frac{1}{n} \cdot \sum_{j=1}^{n} [1 - \delta(s_{j1}, s_{j2})],$$

where s_{j1} and s_{j2} are the states of individuals i_1 and i_2, respectively, at locus f_j. The simple match coefficient $M(i_1, i_2)$ ranges between 0 and 1. The corresponding measure of dissimilarity between individuals is the simple mismatch coefficient for multistate characters:

$$M(i_1, i_2) = 1 - M(i_1, i_2) = \frac{1}{n} \cdot \sum_{j=1}^{n} \delta(s_{j1}, s_{j2})$$

We assume that only one band is produced for each allele by codominant markers, and different alleles generate bands at distinct positions. For each individual its vector of states $i = (s_1, s_2, \dots, s_n)$ can be encoded into a binary pattern on

$$n_b = \sum_{j=1}^{n} r_j$$

binary factors (total number of alleles for all loci). The set $S_j = \{s_{j1}, s_{j2}, \dots, s_{jr}\}$ of all $r = r_j$ different alleles of locus f_j is recoded into the binary form as follows:

S_j	p_{j1}	p_{j2}	p_{j3}	…	p_{jr}
s_{j1}	1	0	0	…	0
s_{j2}	0	1	0	…	0
s_{j3}	0	0	1	…	0
…	…	…	…	…	…
s_{jr}	0	0	0	…	1

where $p_{j1}, p_{j2}, \dots, p_{jr}$ are the $r = r_j$ binary factors (band positions) corresponding to the alleles from multiallelic locus f_j ($j = 1, 2, \dots, n$). Then two individuals i_1 and i_2 can be compared by means of their binary patterns obtained according to the recoding procedure. These binary patterns are (0,1)-vectors, where 1 and 0 correspond to presence and absence, respectively, of a band at some position. The size of this vector equals the total number of possible band positions, and it is in fact the total number of alleles n_b because of the one-to-one correspondence between bands and alleles.

The number a of positions with shared bands (1s) for the both individuals, the number b of positions where individual i_1 has a band, but i_2 does not, and the number c of factors where individual i_2 has a band but i_1 does not, may be expressed as follows:

$$a = \sum_{j=1}^{n} [1 - \delta(s_{j1}, s_{j2})], \quad \text{and} \quad b = c = \sum_{j=1}^{n} \delta(s_{j1}, s_{j2})$$

Then the Dice coefficient of similarity between two individuals i_1 and i_2 has the following form:

$$\begin{aligned} D(i_1, i_2) &= \frac{2a}{2a+b+c} = \frac{2 \cdot \sum_{j=1}^{n} [1-\delta(s_{j1}, s_{j2})]}{2 \cdot \sum_{j=1}^{n} [1-\delta(s_{j1}, s_{j2})] + \sum_{j=1}^{n} \delta(s_{j1}, s_{j2}) + \sum_{j=1}^{n} \delta(s_{j1}, s_{j2})} \\ &= \frac{\sum_{j=1}^{n} [1-\delta(s_{j1}, s_{j2})]}{n} \\ &= M(i_1, i_2) \end{aligned} \tag{1}$$

Therefore, the values of the simple match coefficient for multiallelic loci (multistate characters) equal those of the Dice coefficient of similarity between the corresponding banding profiles (binary patterns) of haploid individuals.

Diploid (polyploid) organisms

Let us consider a diploid individual i subject to genetic analysis by codominant molecular markers in n loci. We denote $r = r_j$ to be the number of alleles in a multiallelic locus f_j ($j = 1, 2, \dots, n$), and $S_j = \{s_{j1}, s_{j2}, \dots, s_{jr}\}$ is the set of these alleles. Then an individual i can be represented by its vector of states $i = \{t_1, t_2, \dots, t_n\}$, where t_j is one of the homozygous $t_j = < s_{ju}s_{ju} >$ or heterozygous $t_j = < s_{ju}s_{jv} >$ states of locus f_j ($u, v = 1, 2, \dots, r$; $j = 1, 2, \dots, n$). If two individuals i_1 and i_2 are represented by their state-patterns $i_k = (t_{1k}, t_{2k}, \dots, t_{nk})$ for $k = 1, 2$, then the question is how to measure similarity between them. There are two steps that should be considered: (i) assessing similarity of homozygous and heterozygous states within the same locus, and (ii) measuring similarity between individuals across all loci on the basis of within locus comparisons.

The only mathematical method for measuring genetic distance between diploid organisms with unknown pedigree information for codominant markers was developed by Peakall *et al.* (1995) and extensively explained in Smouse & Peakall (1999). To obtain a multilocus distance, they logically add the values of dissimilarities scored within each locus. However, the approach they proposed for assessment of dissimilarity with respect to a single locus for multiallelic diploid genotypes looks rather mechanistic (geometric) and does not have any genetic basis [at least no justification was presented in Smouse & Peakall (1999)]. The Euclidean metric in (r-1)-dimensional space was used for measuring dissimilarity between homozygotes and heterozygotes generated by r alleles, where r homozygotes are represented by the vertices of an equilateral (r-1)-dimensional pyramid (interval for $r = 2$, triangle for $r = 3$, tetrahedron $r = 4$, etc.), distance between these vertices (homozygotes) equals 2, and $r(r-1)/2$ heterozygotes are positioned midway between the respective homozygotes. For example, in the case of three alleles A, B and C the geometric distances between two homozygotes AA and BB, and between homozygote AA and heterozygote BC equal

2 and $\sqrt{3}$, respectively. We cannot find any genetic reason why two homozygotes *AA* and *BB* should be more dissimilar than homozygote *AA* and heterozygote *BC*.

In our approach for assessing similarity between homozygotes and heterozygotes, we assume that for four alleles *A*, *B*, *C* and *D* identity-in-state between two individuals can be defined by letting *AA* and *AA*, and *AB* and *AB* comparisons indicate absolute (100%) identity, by letting *AA* and *AB*, and *AB* and *AC* comparisons indicate 50% identity, and letting *AA* and *BB*, *AA* and *BC*, and *AB* and *CD* comparisons indicate absolute dissimilarity (0% identity). This means that unlike Smouse & Peakall (1999), we consider two pairs of genotypes *AA* and *BB*, and *AA* and *BC* as equally dissimilar.

Formally, we assess similarity of homozygous and heterozygous states within the same locus according to the following algorithm. Firstly, for two alleles s_u and s_v of locus *f* we define δ-distance $\delta(s_u, s_v) = 0$ if $u = v$, and $\delta(s_u, s_v) = 1$ if $u \neq v$ $(u, v = 1, 2, \dots, r)$. If A_l designates one of the alleles of locus *f*, then for *q*-ploid organisms all homozygous and heterozygous states $t_A = < A_1A_2 \dots A_q >$ within locus *f* are determined up to permutations of alleles $A_1, A_2, \dots, A_q$ (A_i and A_j do not necessary represent different alleles). For instance, in the case of tetraploid the homozygotes or heterozygotes $< A_1A_2A_3A_4 >$, $< A_2A_4A_3A_1 >$, $< A_4A_3A_2A_1 >$, etc. are genetically identical. Therefore secondly, distance between two homozygous or heterozygous states, $t_A = < A_1A_2 \dots A_q >$ and $t_B = < B_1B_2 \dots B_q >$, is defined as follows. To each allele A_i from one genotype, an allele B_j from the second genotype is matched so as (i) to generate *q* different pairs of alleles where all alleles A_i and B_j are involved and each allele appears in just one pair, and (ii) to minimize the sum of δ-distances $\delta(A_i, B_j)$ between *q* corresponding pairs of alleles [$\delta(A_i, B_j) = 1$ if $A_i = B_j = s_u$, and $\delta(A_i, B_j) = 0$ if $A_i = s_u$, $B_j = s_v$ and $u \neq v$]. There are

$$q! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot q$$

possibilities of the matching between alleles (for instance, for tetraploid $q = 4$ and $q! = 24$). Finding the best matches is known as the 'assignment problem' in operation research (Bellman *et al.* 1970). The distance between two states t_A and t_B within the locus is determined as the minimum sum of δ-distances $As(t_A, t_B)$ derived for the best matches. It is possible to normalize this distance, so that it ranges from 0 to 1, by dividing the obtained minimum value of the sum of δ-distances between matched pairs of alleles by the number of chromosomes *q*. We consider the normalized version of $As(t_A, t_B)$

$$as(t_A, t_B) = \frac{As(t_A, t_B)}{q} \tag{2}$$

as the measure of dissimilarity between two genetic states within the same locus. Then the corresponding measure of similarity is $1 - as(t_A, t_B)$. Thus, the idea of measuring dissimilarity between two homozygotes or heterozygotes within a locus is to match up alleles of the two individuals so that there are as few discordant alleles (across pairs) as possible. Results of calculations of genetic dissimilarity within one locus in the case of three alleles for diploid and two alleles for tetraploid forms are presented in Tables 1 and 2, respectively.

Table 1 Genetic dissimilarity within one locus in the case of three alleles for diploid forms [according to formula (2) — above diagonal; according to Smouse & Peakall (1999) — below diagonal]

	AA	*BB*	*CC*	*AB**	*AC**	*BC**
AA	0	1	1	1/2	1/2	1
BB	2	0	1	1/2	1	1/2
CC	2	2	0	1	1/2	1/2
AB	1	1	$\sqrt{3}$	0	1/2	1/2
AC	1	$\sqrt{3}$	1	1	0	1/2
BC	$\sqrt{3}$	1	1	1	1	0

*Heterozygous states *BA*, *CA*, and *CB* are identical to *AB*, *AC*, and *BC*, respectively.

The following example shows why direct application of either the Dice (*d*), or Jaccard (*j*), or simple mismatch (*m*) coefficients to the banding patterns of codominant locus for diploid genotypes is inappropriate. Let *A*, *B*, *C* and *D* be four alleles, each of which is represented by a single specific band at some position and coded by four-dimensional (0,1)-vectors (1 and 0 correspond to presence and absence, respectively, of a band) as follows: *A* = (1000), *B* = (0100), *C* = (0010), *D* = (0001). Then the same vector like the corresponding allele represents four homozygous states, whereas six heterozygous states are represented by vectors with two 1 s at the positions corresponding to the alleles involved:

AA	*BB*	*CC*	*DD*	*AB*	*AC*	*AD*	*BC*	*BD*	*CD*
1	0	0	0	1	1	1	0	0	0
0	1	0	0	1	0	0	1	1	0
0	0	1	0	0	1	0	1	0	1
0	0	0	1	0	0	1	0	1	1

Then the dissimilarity between genetic states *AA* and *AB* equals *as*(*AA*, *AB*) = 1/2, and it is different from the values of the Dice and simple mismatch coefficients of dissimilarity between the corresponding binary patterns: *d*[(1000), (1100)] = 1/3 and *m*[(1000), (1100)] = 1/4, respectively. On the other hand, the dissimilarity *as* between heterozygotes *AB* and *AC* also equals 1/2, and it is different from the value of the Jaccard coefficients of dissimilarity between the corresponding binary patterns, *j*[(1100), (1010)] = 2/3. Moreover, the pairs of genetic states *AA* and *AB*, and *AB* and *AC* are not equally distant according to the three mentioned coefficients 1/3 = *d*[(1000), (1100)] ≠ *d*[(1100), (1010)] = 1/2, 1/2 = *j*[(1000), (1100)] ≠ *j*[(1100), (1010)] = 2/3,

Table 2 Genetic dissimilarity within one locus in the case of two alleles for tetraploid forms [according to formula (2)]

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	*AAAA*	0															
2	*AAAB*	1/4	0														
3	*AABA*	1/4	0	0													
4	*ABAA*	1/4	0	0	0												
5	*BAAA*	1/4	0	0	0	0											
6	*AABB*	1/2	1/4	1/4	1/4	1/4	0										
7	*ABAB*	1/2	1/4	1/4	1/4	1/4	0	0									
8	*ABBA*	1/2	1/4	1/4	1/4	1/4	0	0	0								
9	*BAAB*	1/2	1/4	1/4	1/4	1/4	0	0	0	0							
10	*BABA*	1/2	1/4	1/4	1/4	1/4	0	0	0	0	0						
11	*BBAA*	1/2	1/4	1/4	1/4	1/4	0	0	0	0	0	0					
12	*ABBB*	3/4	1/2	1/2	1/2	1/2	1/4	1/4	1/4	1/4	1/4	1/4	0				
13	*BABB*	3/4	1/2	1/2	1/2	1/2	1/4	1/4	1/4	1/4	1/4	1/4	0	0			
14	*BBAB*	3/4	1/2	1/2	1/2	1/2	1/4	1/4	1/4	1/4	1/4	1/4	0	0	0		
15	*BBBA*	3/4	1/2	1/2	1/2	1/2	1/4	1/4	1/4	1/4	1/4	1/4	0	0	0	0	
16	*BBBB*	1	3/4	3/4	3/4	3/4	1/2	1/2	1/2	1/2	1/2	1/2	1/4	1/4	1/4	1/4	0

and $1/4 = m[(1000), (1100)] \neq m[(1100), (1010)] = 1/2$. Therefore, none of the commonly used measures of dissimilarity between binary patterns adequately assesses genetic dissimilarity *as* between homozygous and heterozygous states. By the way, this statement holds true for geometric distance of Smouse & Peakall (1999) as well, because by Smouse and Peakall's measure, the genetic states *AA* and *AB*, and *AB* and *AC* are also equally distant (dissimilarity equals 1).

We propose the following relevant approach to assess dissimilarity between diploid individuals based on codominant molecular markers analyses.

1 Establish the number of alleles r_j within each multiallelic locus f_j $(j = 1, 2, \dots, n)$ for n codominant loci. The number of alleles equals the total number of different bands obtained within the locus for all individuals studied.

2 Transform the (0,1)-vector representing the banding pattern of each individual i in locus f_j to the corresponding homozygous $t_j = < s_{ju}s_{ju} >$ or heterozygous state $t_j = < s_{ju}s_{jv} >$, where s_{ju} and s_{jv} $(u, v = 1, 2, \dots, r_j)$ are the matching alleles of locus f_j. Then individual i is represented by its vector of states $i = (t_1, t_2, \dots, t_n)$.

3 Calculate according to formula (2) values of genetic dissimilarity between two individuals $i_1 = (t_{11}, t_{21}, \dots, t_{n1})$ and $i_2 = (t_{12}, t_{22}, \dots, t_{n2})$ with respect to each locus f_j. The values of $as(t_{j1}, t_{j2})$ for $j = 1, 2, \dots, n$ are derived for the corresponding genetic states of individuals within each of n loci.

4 Assess genetic dissimilarity between two individuals i_1 and i_2 by averaging within locus genetic dissimilarities across all n loci:

$$as(i_1, i_2) = \frac{1}{n} \cdot \sum_{j=1}^{n} as(t_{j1}, t_{j2}) \tag{3}$$

This measure varies between 0 for two 'identical' individuals and 1 for two individuals that do not share any allele (band) across all codominant loci tested.

Fingerprint profiles of polyploid organisms with codominant markers do not generally allow distinguishing between two different heterozygous states. For example, if three alleles *A*, *B* and *D* have the following banding patterns *A* = (1000), *B* = (0100) and *D* = (0001), then heterozygous states < *AABD* >, < *ABBD* > and < *ABDD* > of tetraploid individuals have the same codominant profile (1101). This means that index (3) cannot be used directly for measuring dissimilarity between polyploid individuals. Banding patterns of polyploid organisms with codominant markers may not express individuals' genotypes and should be considered only as phenotypes. Comparison between phenotypes can be realized with any measure of dissimilarity, however, studies that employ different measures of similarity should not be assumed to give consistent results.

Dominant markers and similarity measures

For dominant markers such as RAPDs, AFLPs and ISSRs, it is generally assumed that each band represents a different locus and that the alternative to a band at the gel position characteristic of that locus is the absence of a band anywhere in the gel. Thus, for dominant markers there is a direct identity assumed between the number of unique bands observed and the number of identifiable loci for the sample of individuals. On the other hand, the interpretation of shared absences of specific bands by two individuals may depend on the degree of genetic similarity among individuals within the sample. That is, the interpretation may be different when the individuals are drawn from different taxa in a phylogenetic tree than when the individuals are all from closely related populations of a single species.

Haploid organisms

Data obtained for haploid organisms with dominant markers (RAPD, AFLP and ISSR) present a greater challenge than codominant markers to the choice of the best similarity coefficient to represent phenotypic or genetic similarities between pairs of isolates. The main problem here is treatment of the shared absence of a band at some position by two individuals. A common argument against using the simple match coefficient for dominant marker data is that the shared absence of a band by two individuals should not be regarded as evidence of similarity between the two individuals. The commonly stated basis of the argument is that the absence of a trait may result from many different causes and therefore the shared absence of any trait is not good evidence of genetic similarity. This argument, however, ignores the high degree of DNA sequence identity among members of a single species of fungi or other eukaryotes. For example, Birren *et al*. (2003) reported an estimated sequence divergence between two isolates of the fungus, *Saccharomyces cerevisiae* of independent origin to be only 0.5–1.0%, which is slightly less than the sequence divergence reported between humans and chimpanzees. It is common in AFLP analysis of fungal isolates for the presence or absence of bands to be determined by sequences of seven and eight base pairs at either end of the DNA fragment that is amplified by the primers used. Thus, when two isolates exhibit a band at a given position, it means that they have identical DNA sequences over the combined 15 base pair regions of their two genomes. On the other hand, even in the absence of a band at that position for two other isolates of the same fungal species, we can estimate assuming independence of nucleotides in DNA sequences that the two isolates have a probability of identical sequences over the 15 base pair region of between $(0.99)^{15} \sim 0.86$ and $(0.995)^{15} \sim 0.93$ even if those isolates come from geographically separated populations. Thus, assuming that the two isolates of *S. cerevisiae* are genetically different in the critical DNA sequences because they each lack a specific AFLP band is likely to be wrong between 86 and 93% of the time. A more conservative approach would be to assume, as the simple match coefficient does, that the shared lack of a specific AFLP band by two isolates of the same fungal species is good probable evidence that the two are genetically similar for that trait.

Using shared absences of specific bands as evidence of genetic similarity requires a decision of which band positions should be considered in the comparison of each pair of individuals. Obviously, we cannot identify a band for consideration unless that band is polymorphic in the populations under consideration. If a particular dominant band is found with one individual of a population, we may use that band as a factor to be considered in the comparison of genetic similarity between other individuals of the population even if some pairs share a common absence of that band.

Of course, for comparisons of genetic similarity between individuals of different species, a much lower level of DNA sequence identity would be expected. Kellis *et al*. (2003) estimate the sequence divergence between *Saccharomyces paradoxus* and *S. cerevisiae* to be 20%, which is greater than that between humans and rhesus monkeys. In a comparison between isolates of *S. paradoxus* and *S. cerevisiae*, there would be a probability of sequence identity over 15 base pairs at an AFLP locus of only $(0.8)^{15} \sim 0.04$. In the comparison between these two species, the conservative approach would be to disregard shared absences of specific AFLP bands as potential evidence of similarity. Thus, the simple match coefficient would not be a good choice for comparing phylogenetic relationships between species of *Saccharomyces* based on AFLP data. Landry & Lapoint (1996) suggested that the Dice or Jaccard coefficients might be preferable to the simple match coefficient when using RAPD analysis to compare groups of distantly related taxa. Hallden *et al.* (1994) considered the simple match coefficient to be the more appropriate measure of similarity when closely related taxa are considered, but we believe that choice should be supported with estimates of DNA sequence identity between the taxa. In the absence of supporting sequence identity estimates, similarity values based on dominant markers data should be regarded as tentative.

Thus, the simple match coefficient is the most suitable measure of similarity in the case of closely related haploid individuals when very low DNA sequences divergence between two individuals is expected. Otherwise, when rather distinct individuals are compared and the shared absence of a band does not contribute to similarity of individuals, the Jaccard coefficient of similarity is preferable to the Dice coefficient. We could not find any justification for giving more weight to the shared presence of bands when comparing individuals from different taxa. Therefore, we prefer the Jaccard coefficient over the Dice coefficient, or the mechanistic application of the Nei & Li (1979) method.

Diploid (polyploid) organisms

The problem with dominant markers for diploids is that, without genetic data from segregation patterns after selfing, it would be impossible to distinguish bands that represent two alleles at a homozygous locus from bands that represent only one allele at a heterozygous locus. Thus, it would be generally impossible to determine exact genetic similarity between two individuals that share a band at the same position. Therefore, similarity values based on dominant markers with diploids should be regarded only as rough estimates that are based on incomplete information.

Estimates of integral parameters of populations or groups of individuals (for example, allele frequencies, average dissimilarity between individuals, diversity of populations, etc.) could be improved for those cases in which it is realistic to assume that random mating occurs in the diploid species and genotype frequencies follow the proportions of HWE. That might allow calculation of the probable level of heterozygosity at the locus represented by each band based on its frequency in the entire sample of isolates tested (Lynch & Milligan 1994). This is based on the assumption of just two alleles per locus: one that gives a band at the observed position and one that gives no band anywhere in the gel. Departure from Hardy–Weinberg proportions on the estimates of alleles frequencies was analysed by Zhivotovsky (1999).

Unfortunately, the HWE-based approach does not allow improving estimation of similarity between any two specific individuals. For diploid organisms that are primarily inbreeders, we would expect a low level of heterozygosity, so it might be sufficient to treat those organisms as if they were haploids. For diploid organisms that reproduce asexually there might not be any good way to estimate similarity with dominant markers, so a rough estimate based on phenotypic similarity might be all that is possible.

As no suitable method can be proposed for measuring genetic similarity between diploid organisms on the basis of dominant banding profiles, we cannot recommend any preferred similarity measure for dominant markers in diploid (polyploid) organisms. On the other hand, any index might be used for phenotypic comparison between fingerprint profiles considered rather as phenotypes than genotypes. For reasons described earlier in regard to haploids, we prefer the simple match coefficient for comparisons of phenotypic similarity in populations within a single diploid or polyploid species.

Conclusions

The principal problem with analysis of genetic relationships between individuals with molecular markers is measuring their dissimilarity. There are no acceptable universal approaches for assessing genetic dissimilarity between individuals based on molecular markers. Different dissimilarity measures are relevant to, and should be used with, multilocus dominant and codominant DNA markers as well as with diploid (polyploid) and haploid individuals.

The Dice dissimilarity index is the suitable for haploids with codominant molecular markers, and it can be applied directly to (0,1)-vectors representing multilocus multiallelic banding profiles of individuals. None of the Dice, Jaccard and simple mismatch coefficient is appropriate for diploids (polyploids) with codominant markers, because there is no way for direct processing of fingerprint profiles. By transforming multiallelic banding patterns at each locus into the corresponding homozygous or heterozygous states, a new measure of dissimilarity within loci was developed (formula 2) and expanded for measuring dissimilarity between multilocus states of two individuals by averaging across all codominant loci tested (formula 3).

The simple mismatch coefficient can be considered as the most suitable measure of dissimilarity between banding patterns of closely related haploid forms, whereas for distantly related haploid individuals the Jaccard dissimilarity is recommended. In general, no suitable method for measuring genetic dissimilarity between diploids with dominant markers can be proposed. Therefore, analyses of genetic dissimilarity between diploid (polyploid) organisms with dominant markers should be viewed with caution unless the organism is highly inbred and therefore highly homozygous. Banding patterns of polyploids with codominant markers and diploids with dominant markers represent individuals' phenotypes rather than genotypes. Descriptive comparison of phenotypes with different indices is possible and relevant, but any genetic inferences cannot be justified in such a case.

All dissimilarity measures proposed and developed herein are metrics.

Acknowledgements

We gratefully acknowledge financial support of the Colton Foundation (Tel Aviv University) to E. Kosman.

References

Anthony F, Combes MC, Astorga C, Bertrand B, Graziosi G, Lashermes P (2002) The origin of cultivated *Coffea arabica* L. varieties revealed by AFLP and SSR markers. *Theoretical and Applied Genetics*, **104**, 894–900.

Barth S, Melchinger AE, Lübberstedt TH (2002) Genetic diversity in *Arabidopsis thaliana* L. Heynh. investigated by cleaved amplified polymorphic sequence (CAPS) and inter-simple sequence repeat (ISSR) markers. *Molecular Ecology*, **11**, 495–505.

Belaj A, Satovic Z, Cipriani G *et al*. (2003) Comparative study of the discriminating capacity of RAPD, AFLP and SSR markers and of their effectiveness in establishing genetic relationships in olive. *Theoretical and Applied Genetics*, **107**, 736–744.

Bellman R, Cooke KL, Lockett JA (1970) *Algorithms, Graphs and Computers*. Academic Press, New York.

Birren B, Fink G, Lander E (2003) Fungal genome initiative: a white paper for fungal comparative genomics. *The Fungal Genome Initiative, Second White Paper*. www.genome.wi.mit.edu/annotation/fungi/fgil.

Bleeker W (2003) Hybridization and *Rorippa austriaca* (Brassicaceae) invasion in Germany. *Molecular Ecology*, **12**, 1831–1841.

Brouat C, Mckey D, Douzery EJP (2004) Differentiation in a geographical mosaic of plants coevolving with ants: phylogeny of the *Leonardoxa africana* complex (Fabaceae: Caesalpinioideae) using amplified fragment length polymorphism markers. *Molecular Ecology*, **13**, 1157–1171.

Cordeiro GM, Pan YB, Henry RJ (2003) Sugarcane microsatellites for the assessment of genetic diversity in sugarcane germplasm. *Plant Sciences*, **165**, 181–189.

Dearborn DC, Anders AD, Schreiber EA, Adams RMM, Mueller UG (2003) Inter-island movements and population differentiation in a pelagic seabird. *Molecular Ecology*, **12**, 2835–2843.

Hallden C, Nilsson NO, Rading IM, Sall T (1994) Evaluation of RFLP and RAPD markers in a comparison of *Brassica napus* breeding lines. *Theoretical and Applied Genetics*, **88**, 123–128.

Huff DR, Peakall R, Smouse PE (1993) RAPD variation within and among natural populations of outcrossing buffalograss [*Buchloe dactyloides* (Nutt.) Engelm.]. *Theoretical and Applied Genetics*, **86**, 927–934.

Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, **423**, 241–254.

Kosman E (2003) Nei's gene diversity and the index of average differences are identical measures of diversity within populations. *Plant Pathology*, **52**, 533–535.

Kumar J, Nelson RJ, Zeigler RS (1999) Population structure and dynamics of *Magnaporthe grisea* in the Indian Himalayas. *Genetics*, **152**, 971–984.

Landry PA, Lapointe FJ (1996) RAPD problems in phylogenetics. *Zoologica Scripta*, **25**, 283–290.

Lynch M, Milligan BG (1994) Analysis of population genetic structure with RAPD markers. *Molecular Ecology*, **3**, 91–99.

Maguire TL, Peakall R, Saenger P (2002) Comparative analysis of genetic diversity in the mangrove species *Avicennia marina* (Forsk.) Vierh. (Avicenniaceae) detected by AFLPs and SSRs. *Theoretical and Applied Genetics*, **104**, 388–398.

Mock KE, Brim-Box JC, Miller MP, Downing ME, Hoeh WR (2004) Genetic diversity and divergence among freshwater mussel (*Anodonta*) populations in the Bonneville basin of Utah. *Molecular Ecology*, **13**, 1085–1088.

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleasis. *Proceedings of the National Academy of Sciences of the USA*, **76**, 5269–5273.

Pannebakker BA, Zwaan BJ, Beukeboom LW, Van Alphen JJM (2004) Genetic diversity and *Wolbachia* infection of the *Drosophila* parasitoid *Leprophilina clavepes* in western Europe. *Molecular Ecology*, **13**, 1119–1128.

Peakall R, Smouse PE, Huff DR (1995) Evolutionary implications of allozyme and RAPD variation in diploid populations of dioecious buffalograss *Buchloë dactyloides*. *Molecular Ecology*, **4**, 135–147.

Pei MH, Ruiz C (2000) AFLP evidence of the distinctive patterns of life-cycle in two forms of *Melampsora rust* on *Salix viminalis*. *Mycological Research*, **104**, 937–942.

Schönswetter P, Paun O, Tribsch A, Niklfeld H (2003) Out of the Alps: colonization of Northern Europe by East Alpine populations of the Glacier Buttercup *Ranunculus glacialis* L. (Ranunculaceae). *Molecular Ecology*, **12**, 3373–3381.

Smouse PE, Peakall R (1999) Spatial autocorrelation analysis of individual multiallele and multilocus genetic structure. *Heredity*, **82**, 561–573.

Sneath PA, Sokal RR (1973) *Numerical Taxonomy*. W.H. Freeman Co, San Francisco.

Teulat B, Aldam C, Trehin R *et al*. (2000) An analysis of genetic diversity in coconut (*Cocos nucifera*) populations from across the geographic range using sequence-tagged microsalellites (SSRs) and AFLPs. *Theoretical and Applied Genetics*, **100**, 764–771.

Torimaru T, Tomaru N, Nishimura N, Yamamoto S (2003) Clonal diversity and genetic differentiation in *llex leucoclada* M. patches in an old-growth beech forest. *Molecular Ecology*, **12**, 809–818.

Zhivotovsky LA (1999) Estimating population structure in diploids with multilocus dominant DNA markers. *Molecular Ecology*, **8**, 907–913.

Dr Evsey Kosman (PhD in mathematics) and Prof Kurt Leonard have common research interests in analyses of population structure, linkage disequilibrium and diversity within and between plant pathogen populations. Kosman also works on development more general approaches for measuring diversity and data analysis, and Leonard works on population genetics of plant pathogenic fungi.