CONCEPTUAL AND STATISTICAL ISSUES RELATED TO THE USE OF MOLECULAR MARKERS FOR DISTINCTNESS AND ESSENTIAL DERIVATION

PROBLEMES CONCEPTUELS ET STATISTIQUES LIES A L'UTILISATION DE MARQUEURS MOLECULAIRES POUR LA DISTINCTION ET LA DERIVATION ESSENTIELLE

F.A. van Eeuwijk[1] and C.P. Baril[2]
[1]Wageningen University, Dept of Plant Sciences, Laboratory of Plant Breeding, P.O. Box 386, 6700 AJ Wageningen, The Netherlands
[1]Plant Research International, Dept. of Biodiversity and Identity, P.O. Box 16, 6700 AA Wageningen, The Netherlands
[2]GEVES, La Minière, 78285 Guyancourt Cedex, France

Abstract

An exhibition is given of old and new statistical procedures for dealing with marker information in the context of distinctness testing and assessing genetic conformity for essential derivation purposes. Conceptual issues are discussed in relation to statistical methods. It is believed that the most important statistical and conceptual difference between distinctness and conformity testing resides in the wording of null and alternative hypotheses. For distinctness testing, the null hypothesis states no difference between varieties, while the alternative implies the existence of a difference. For conformity testing, null and alternative hypothesis are non-equivalence and equivalence, respectively. The reversal of null and alternative hypothesis has rather limited statistical consequences when test statistics are distance measures. Characteristically, morphological characters form the preferred traits for assessing distinctness, while molecular markers are chosen for assessing conformity. From a statistical point of view this difference is rather immaterial. Distinctness and conformity are throughout presented as two closely related concepts, whose assessment takes place by highly comparable statistical procedures. Specific topics that are addressed in the paper are first the present positions of UPOV and ASSINSEL. Subsequently, uni- and multivariate methods for distinctness and conformity are treated, separately for genetically homogeneous and heterogeneous varieties. Lastly, the choice of markers is discussed, as are the relations between morphological, marker and pedigree information.

Keywords: bootstrap, distinctness, essential derivation, genetic distance, genetic similarity, plant variety protection, permutation test

1. Introduction

Opportunities provided by molecular markers form the focus of discussion in affairs concerning the granting and limiting of intellectual property rights for new plant varieties. Key players in the discussion are UPOV (The International Union for the Protection of New Varieties of Plants; http://www.upov.int/eng/index.htm) and ASSINSEL (The International Association of Plant Breeders for the Protection of Plant Varieties; http://www.worldseed.org/assinsel.htm). UPOV is considering, or forced to consider, the role that molecular markers can play in the granting of plant breeders' rights. Presently, breeders' rights are granted to a candidate variety on the basis of species specific tests as defined in UPOV guidelines. A candidate should comply with the requirements of novelty, distinctness, uniformity and stability, the latter three forming the components of the wellknown DUS procedure. Distinctness is the predominant requirement, uniformity and stability act as subsidiary requirements. The characters for

Proc. Int. Symp. on Molecular Markers
Eds. Doré, Dosba & Baril
Acta Hort. 546, ISHS 2001

the DUS procedure are primarily morphological in nature. For allogamous crops, distinctness is assessed by a t-test comparing the variety mean of a candidate with the means of the varieties of a reference collection (and other candidate varieties). Distinctness is conferred when the candidate is significantly different from all other varieties at least one trait. The additional requirement of uniformity asks a candidate variety not to be more variable than reference varieties that are comparable in mean. The further requirement of stability applies when a variety retains its characteristics through repeated cycles of propagation.

With the steep increase in popularity of molecular markers the temptation has arisen to introduce markers as characters for distinctness. This choice would make available a great number of new characteristics (markers), that are in principle easily observable and free of genotype by environment interaction. Reservations exist due to the worry that molecular markers for the greater part bear an unclear relation to the phenotype in general, and that acceptance of markers will lead to erosion of plant variety protection because of the gradual decrease of the minimum distance between varieties that is currently implicit in the determination of distinctness.

Where UPOV's principal preoccupation is located in the granting of plant breeders' rights, ASSINSEL's involvement concerns first and foremost the range of applicability of these rights. ASSINSEL wants a procedure that would protect its members from infringement and piracy on their genetic material. Roughly stated, the issue is to withhold breeders' rights from a certain type of 'new' variety that has been derived directly from an already protected variety, with the aim of copying as much as possible of the protected variety. The 'new' variety is highly genetically similar, because of the explicit use of the protected variety in the breeding process, but simultaneously just enough phenotypic difference was introduced to allow the 'new' variety to pass the UPOV distinctness test, which would make the 'new' variety eligible for breeders' rights. The older, protected variety is called the initial variety, the variety derived from it is called essentially derived when some requirements are complied with. The most controversial of these requirements concerns the estimation of the degree of conformity between initial and potentially derived variety. Dispute exists around empirical and statistical questions regarding the optimal traits and test statistic to be used. Originally ASSINSEL exhibited a strong predilection for the use of molecular markers to estimate genetic conformity on the familiar grounds of quick and relatively cheap use, and absence of genotype by environment interaction. Recently, this position has slightly changed so as to allow expression related traits to codetermine conformity.

Discussions on the use of markers for distinctness have so far been rather heated, because of the insecurity about the consequences of such a move for the minimum distance needed for distinctness and the interpretation of what such a distance would mean. Discussions on the assessment of essential derivation suffered from confusion about how genetic conformity should be measured. We believe that a better insight in the statistical aspects related to how to use marker information for distinctness and essential derivation will contribute to a more transparent discussion on the utility of markers for plant proprietary issues. The major objective of this paper is to present statistical methods for assessing distinctness and conformity, with special attention for how to include marker information. Necessarily, also conceptual issues transcending purely statistical arguments will be touched upon. We do not have the pretension of being exhaustive in the sense of treating all questions with regard to the assessment of distinctness and conformity for any kind of crop.

We think that in a statistical sense the major difference between distinctness and conformity (essential derivation) is the phrasing of null- and alternative hypothesis. For distinctness we take equality of varieties as null hypothesis and conclude upon distinctness at rejection of the null hypothesis. On the contrary, for conformity related to essential derivation we assume that the varieties are different to a certain extent under the null hypothesis, whereas rejection of the null hypothesis will lead to equivalence of the varieties. The inversion of null and alternative hypothesis for conformity testing has as a

consequence that statistical procedures for conformity will as a rule be more complicated than those for distinctness. For statistical tests it is necessary to know the behaviour of the test statistic under the null hypothesis. When the null hypothesis states that certain differences will be present between the treatments (varieties), the derivation of the null distribution will cost more effort than when the null hypothesis has the classical no-difference set-up.

The topics in the rest of the paper will be as follows. In section 2 we will give a brief state of the art overview of distinctness and essential derivation. In section 3 statistical procedures for assessing distinctness and conformity in crops with genetically heterogeneous varieties will be presented. This will be a mixture of old and new methods. Section 4 will treat procedures for crops with genetically homogeneous varieties. Finally, section 5 will contain a number of discussion points.

2. Distinctness and essential derivation, actual situation

2.1. Distinctness

Distinctness is assessed on the basis of the expression of morphological, physiological or even biochemical characteristics, proposed in the guidelines of UPOV. Phenotypic expression is observed over at least two independent cycles (years and/or sites), with the exception of many ornamental species, which are reproduced by vegetative propagation, what requires only one cycle, usually in a greenhouse. Depending on the species, UPOV guidelines define between 15 and to 50 traits to be observed on each variety. The traits are grouped into compulsory traits, recommended traits, and complementary traits. Biochemical characteristics belong to the third class and can only provide supporting evidence for distinctness. Each country is free to use additional traits, which usually have no agronomical interest, if they satisfy the criteria for distinctness traits. If molecular markers are to appear at all in the list of traits for distinctness, they are most prone to appear as complementary traits. For genetically heterogeneous varieties the recommended procedure for decisions on distinctness is the Combination Over Years Distinctness (COY-D) test (UPOV, 1997b). UPOV document TWC/15/6 (UPOV, 1997a) shows that COY-D is used by a number of European countries for assessing distinctness in cross pollinating and partially cross pollinating species, among which grasses, clovers, beets and ornamentals. t-tests are used for the comparison of the mean of a candidate variety with the mean of all reference varieties and other candidate varieties. The null hypothesis for the comparison of two varieties g and g' is H_0: $\mu_g = \mu_{g'}$, or, stated differently, $H_0 : \mu_g - \mu_{g'} = 0$, which in turn is equivalent to $H_{01} : \mu_g - \mu_{g'} \geq 0$ *and* $H_{02} : \mu_g - \mu_{g'} \leq 0$. The alternative hypothesis is $H_a : \mu_g \neq \mu_{g'}$, which is equivalent to $H_{a1} : \mu_g - \mu_{g'} < 0$ *or* $H_{a2} : \mu_g - \mu_{g'} > 0$. The test statistic is $T = \frac{\overline{x}_g - \overline{x}_{g'}}{SE_{\overline{x}_g - \overline{x}_{g'}}}$, with the standard error for the difference, $SE_{\overline{x}_g - \overline{x}_{g'}}$, being derived from the variety x year mean square from the analysis of variance on the variety by year table of means. This statistic should follow a t-distribution with degrees of freedom, df, equal to those of the variety x year interaction. The null hypothesis is rejected when $|T| > t_{1-1/2\alpha,df}$, in which $|T|$ is the absolute value of the test statistic and $t_{1-1/2\alpha,df}$ is the 1-1/2α quantile of a t_{df} distribution. UPOV prescribes the test level, α (significance level), for each crop and trait individually. Most test levels are chosen at 1%.

Proc. Int. Symp. on Molecular Markers
Eds. Doré, Dosba & Baril
Acta Hort. 546, ISHS 2001

The COY-D test is essentially a univariate test (a test for one character at a time), without obvious complications as long as the characters are quantitative. To be distinct, a candidate should be found significantly different from all reference varieties in at least one trait.

The present COY-D test can easily be generalized as to include a minimum distance requirement. Let θ be the minimum distance required for a specific character, and define $\theta_L = -\theta$ and $\theta_U = \theta$, then null and alternative hypothesis for distinctness beyond a minimum distance become $H_{01} : \mu_g - \mu_{g'} \geq \theta_L$ *and* $H_{02} : \mu_g - \mu_{g'} \leq \theta_U$. The alternative hypothesis will be $H_{a1} : \mu_g - \mu_{g'} < \theta_L$ *or* $H_{a2} : \mu_g - \mu_{g'} > \theta_U$. We now introduce two test statistics: $T_L = \frac{(\bar{x}_g - \bar{x}_{g'}) - \theta_L}{SE_{\bar{x}_g - \bar{x}_{g'}}}$ and $T_U = \frac{(\bar{x}_g - \bar{x}_{g'}) - \theta_U}{SE_{\bar{x}_g - \bar{x}_{g'}}}$. The null hypothesis of equality below the minimum distance will be rejected at level α when *either* $T_L < t_{1/2\alpha,df}$ *or* $T_U > t_{1-1/2\alpha,df}$.

In contrast to the situation for crops with genetically heterogeneous varieties, for many crops with genetically homogeneous varieties, distinctness is generally determined visually by experts, without having recourse to statistical procedures. For example, in many ornamentals a new flower colour or shape is sufficient for distinctness, provided the character is uniform and stable over the individuals of the candidate variety.

The current situation in the discussion on the possible role of molecular techniques in distinctness testing is that the UPOV Technical Committee has installed 5 subgroups on molecular techniques, more precisely for maize, oilseed rape, rose, tomato and wheat. These subgroups are attached to the UPOV Working Group on Biochemical and Molecular Techniques. The aim of the subgroups is to study and make an inventory of applications of molecular markers that are relevant to the assessment of distinctness, uniformity and stability.

2.2. Essential derivation

ASSINSEL's original intention when for the first time bringing up the concept of essential derivation was to deny breeders' rights to varieties that were obtained by 'cosmetic' breeding. Examples of breeding methods prone to result in essentially derived varieties were given, where in the derivation process explicit use was made of an initial variety. The examples were: selections of natural/ induced mutants, selections of somaclonal variants/ deviating plants, repeated back crosses, transformation and/or genetic engineering.

A recent formulation of the requirements for essential derivation can be found in the consolidated ASSINSEL position paper of the Melbourne conference in 1999. A variety is essentially derived when it shows

- clear distinctness in the sense of the UPOV Convention
- conformity to the initial variety in the expression of the essential characteristics that result from the genotype or combination of genotypes of the initial variety
- predominant derivation from an initial variety.

The distinctness requirement is unproblematic, as it is under the responsibility of UPOV. Predominant derivation can only be decided on by comparison of breeding books, but seems in its definition unproblematic too. The conformity requirement is the one under most discussion within ASSINSEL. The question is how to assess it? The initial conviction to purely base conformity on markers has gradually made place for a position where also phenotypic characteristics and combining ability play a role. To quantify conformity it seems logical to express it in terms of similarities, with 1 meaning complete conformity, or similarity, and 0 meaning complete non-conformity, or dissimilarity.

Alternatively, conformity can be expressed in distance terms, which in turn is often taken to be the complement of similarity. Popular conversions of similarity to distance are distance = 1 - similarity, or distance = $\sqrt{1-\text{similarity}}$. A distance of 0 then means complete conformity, whereas a distance of 1 will indicate complete non-conformity. Conformity, similarity and distance will be used as interchangeable terms from here onwards.

The consequence for a variety of being declared essentially derived is that the breeders' rights remain with the owner of the initial variety, and that authorization of the owner of the initial variety will be necessary for production and marketing of the essentially derived variety.

The actual position of ASSINSEL with regard to the implementation of a conformity criterion mentions the assessment of two thresholds for conformity, to be determined on a species-by-species basis, conditional on the choice of marker system. Below the first threshold a variety should be considered as non-essentially derived from an initial variety. Beyond the second threshold the new variety should be considered as essentially derived, except when the breeder can prove that he used independent germplasm. Between these thresholds there is room for dispute.

Various ASSINSEL project groups study the implementation of essential derivation. Work on maize has resulted in the following conclusions. At the 1998 ASSINSEL conference in Monte Carlo micro-satellites were proposed as molecular technique. Suggested thresholds were put at 85% and 90% 'similarity' (no particular similarity measure seems prescribed). Below 85% non-derivation will be assumed, from 85-90% dispute will follow, above 90% essential derivation will be concluded unless breeding books prove otherwise. Work on tomato ended inconclusive with respect to suggested thresholds. Nevertheless, for some known pedigree relationships, marker based similarities indeed seemed to increase with coancestry. Most progress was booked in ryegrass. A study was started in 1996, comprising 12 diploid perennial ryegrass accessions, under which 5 closely related groups. Morphological and molecular characterizations were carried out (Gilliland *et al.*, 2000, Roldán-Ruiz *et al.*, 2000). The known relationships between the 12 accessions were correctly reflected by AFLPs, and were also consistent with morphological characterizations. A threshold of 7 for the squared Euclidean distance between pairs of varieties was proposed, using 5 specific AFLP primer combinations and a defined DNA protocol. For Euclidean distances below that proposed threshold, new varieties can be considered as possibly essentially derived and various actions can be taken.

Work is continuing within the various ASSINSEL sections. To give one example, a study in lettuce is planned in which the distribution of distance coefficients will be estimated that pertains to initial – essentially derived variety pairs, with the essentially derived varieties constituting the fourth back cross and the initial varieties the recurrent parent. Similar work is in progress for maize, in the sense that also for this crop distributions of distance coefficients are estimated for known initial – essentially derived variety pairs.

Another project that is dedicated exclusively to the development of measuring methodology for genetic conformity, and in which both statistical and molecular aspects will be investigated, is the EU-MMEDV project (QLRT-1999-PL1499; http://www.cordis.lu search MMEDV, or, http://www.niab.com link Research). Three model crops are considered: barley (*Hordeum vulgare*), an autogamous cereal, rose (*Rosa hybrida sp.*) an ornamental crop of high value, and maize (*Zea mays*), an allogamous species.

3. Statistical procedures for assessing distinctness and conformity for genetically heterogeneous varieties

In this section a number of procedures will be described that can be used for assessing distinctness or conformity. Because of the asymmetry between tests for

Proc. Int. Symp. on Molecular Markers
Eds. Doré, Dosba & Baril
Acta Hort. 546, ISHS 2001

distinctness and conformity, some statistical procedures are especially suitable for distinctness testing, others are more suitable for conformity testing, and finally, some are suitable for both distinctness and conformity testing. Although all procedures to be described can in principle be adjusted as to be applicable to both distinctness and conformity testing, the price for such multifunctionality would be loss of transparency. Therefore we proceed with addressing either distinctness or conformity, unless a procedure lends itself without problems for both purposes.

3.1. A univariate procedure using band or allele frequency information

3.1.1. Distinctness

The existing COY-D-test for morphological characters can serve as a mold for a univariate (one trait at a time) statistical procedure for distinctness based on allele frequencies (or band frequencies) for heterogeneous varieties. Two varieties, indexed by g and g', for which profiles are available on n_g and $n_{g'}$ plants, can be compared on the frequency of a particular allele by defining a test-statistic $Z = \frac{(\hat{p}_g - \hat{p}_{g'}) - \theta_0}{SE_{(\hat{p}_g - \hat{p}_{g'})}}$, where $\hat{p}_g$ and $\hat{p}_{g'}$ are the estimates for the population frequencies, p_g and $p_{g'}$, in the varieties g and g'. θ_0 is the difference, or minimum distance, between the population frequencies under the null hypothesis. For classical distinctness tests this parameter is commonly taken as $\theta_0 = 0$, i.e., no difference under the null hypothesis. The estimated standard error for the difference in allele frequencies is $SE_{(\hat{p}_g - \hat{p}_{g'})} = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_g} + \frac{1}{n_{g'}}\right)}$, with $\hat{p} = \frac{n_g \hat{p}_g + n_{g'} \hat{p}_{g'}}{n_g + n_{g'}}$, the average of $\hat{p}_g$ and $\hat{p}_{g'}$. Under the null hypothesis, the statistic Z is approximately a standard normal variable, provided that all of $n_g p$, $n_g(1-p)$, $n_{g'} p$, and $n_{g'}(1-p)$ are equal or greater than 5. The null hypothesis is rejected at level α when $|Z| > z_{(1-1/2\alpha)}$, i.e., when the absolute value of Z is larger than the $1-1/2\alpha$ quantile of the standard normal distribution. A confidence interval approach for testing $p_g = p_{g'}$ rejects the null hypothesis when 0 is outside the interval $((\hat{p}_g - \hat{p}_{g'}) - z_{(1-1/2\alpha)} SE_{(\hat{p}_g - \hat{p}_{g'})}$, $((\hat{p}_g - \hat{p}_{g'}) + z_{(1-1/2\alpha)} SE_{(\hat{p}_g - \hat{p}_{g'})})$. It may be remarked that the procedure shown here for allele frequencies can also serve for the analysis of binary phenotypic traits in standard distinctness tests.

One may doubt whether an approach based on an individual allele will have enough power. For distinguishing a rather small difference between a frequency of 0.45 in a variety g and 0.55 in another variety, g', assuming α=0.05, assessment of distinctness would require more than 100 plants of each variety. In contrast for a big difference of 0.4 between a variety with a frequency of 0.3 and another with 0.7, 12 plants would suffice. Looked at it in this way, a Z-test on individual allele frequencies may prove to be useful.
When information on marker alleles and loci must be combined, a correction should be made for the effects of multiple testing. After all, by considering more loci the probability of rejecting the null hypothesis when actually it is true will increase. As there is no obvious analytic method for combining tests at linked loci, the most appealing option is conferred by a permutation or randomization procedure (Manly, 1997). We illustrate this approach for the distinctness test introduced above, assuming no minimum distance

requirement, $\theta_0 = 0$. The objective is to find the null distribution of Z ($|Z|$) allowing for multiple testing. The general idea behind permutation tests for the comparison of pairs of varieties is that when there exists no real difference between two varieties, it should not matter for the calculation of the test statistic to which of the two varieties individual plants are allocated. The distribution of the test statistic under the null hypothesis (no difference) can be obtained by permuting variety membership a sufficient number of times (e.g., 1000 times), and calculating for each permutation the value of the test statistic. The probability of the original value of the test statistic under the null hypothesis, calculated on the non-permuted data, can be found by comparing this value with the quantiles of the permutation null distribution. An easy way of calculating the tail probability is to sort the original value of the test statistic with the permutation values and assess the extremity of the original value.

For our distinctness test on allele frequencies we have to extend the general principle of permutation testing. To find the null distribution of Z accounting for multiple testing over both linked and unlinked loci and multiple alleles per locus we again permute variety membership for individual plants, but we retain the fingerprints, i.e., the vectors of allele presence/absence. Fingerprints are thus randomly allocated to one of both varieties. Subsequently, for each permutation Z-values are calculated for all loci and alleles, and the maximum (absolute) Z-value over all loci and alleles is stored as an element for the null distribution of the test statistic. Finally, the original $|Z|$ is compared to the quantiles of this null distribution and when it is among the proportion α most extreme values, the null hypothesis of equality of varieties is rejected at test level α.

For establishing a firm criterion it may be good to fix a level α critical value on the basis of the highest critical value found in the permutations for a complete set of reference and candidate varieties. This implies that critical values could be defined once, where agreement on the appropriate set of varieties is required, after that new candidates could be compared on the basis of earlier set critical values.

3.1.2. Conformity

For conformity testing the null hypothesis is $H_{01}: p_g - p_{g'} \leq \theta_L$ *or* $H_{02}: p_g - p_{g'} \geq \theta_U$, i.e., the varieties g and g' are different, while the alternative hypothesis is $H_{a1}: p_g - p_{g'} > \theta_L$ *and* $H_{a2}: p_g - p_{g'} < \theta_U$, or, $H_a: \theta_L < p_g - p_{g'} < \theta_U$, i.e., the varieties are the same within certain limits, where the limits, θ_L and θ_U, are open to discussion. It is not so easy to develop analytically the distribution for the difference in allele frequencies under the null hypothesis, because the null distribution will depend on the unknown real frequencies. Permutation procedures do not provide a solution, as they are easy to work with under null hypotheses of no difference, but become cumbersome for other configurations. The best option seems to construct a bootstrap confidence interval (Efron and Tibshirani, 1993) for a statistic like the maximum absolute difference (or the average absolute difference) in allele frequencies and to test $H_0: \max_{ma} |p_{mag} - p_{mag'}| \geq \theta_{ED}$ versus $H_a: \max_{ma} |p_{mag} - p_{mag'}| < \theta_{ED}$, where the maximum is taken over all loci and alleles within loci.

Bootstrapping is resampling of individuals (plants) with replacement, so that in bootstrap samples some individuals occur more than once, while others are absent. Permutation is resampling without replacement. A bootstrap confidence interval for a parameter as the maximum absolute frequency difference over all loci, $\delta_{g;\,g'} = \max_{ma} |p_{mag} - p_{mag'}|$, is thus created by repeatedly sampling, with replacement, n_g times an individual from variety g, present with n_g individuals, and $n_{g'}$ times an individual from

Proc. Int. Symp. on Molecular Markers
Eds. Doré, Dosba & Baril
Acta Hort. 546, ISHS 2001

variety g', present with $n_{g'}$ individuals. After each sampling, the maximum absolute frequency difference, $\hat{\delta}_{g;\,g'}$ is calculated. Eventually, the quantiles of the bootstrap realizations of the maximum absolute difference are used to define a confidence interval. An appropriate one sided bootstrap confidence interval for the parameter $\delta_{g;\,g'}$ is given by the interval whose upper bound is the 1-α quantile of the bootstrap distribution. If the minimum requirement for non-conformity, θ_{ED}, is beyond the 1-α quantile, then the null hypothesis of the varieties being different is rejected and the varieties are concluded to be conform. The threshold θ_{ED} is subject to discussion. As a rough indication for the minimum number of bootstrap samples we take 2500, from work on inbreeding coefficients by van Dongen and Backeljau (1995).

3.2. A distance based approach for band and allele frequency data

In the former section a correction for multiple testing was constructed by an especially devised permutation procedure for the maximum of the test-statistic over loci and alleles. A more common approach for combining information over alleles and loci is by using distance (similarity) measures. Various choices are possible, but practice learns that estimates of distance measures are usually highly correlated (see for example Sanchez *et al.*, 1995), although differences do exist for the absolute level of the size of distances. For band and allele frequency data, the two most popular classes of distance measures consist of 1) distances based on absolute frequency differences (Manhattan distance) between varieties, and 2) quadratic frequency differences (Euclidean distance). An example of the first class is Prevosti's distance (Prevosti *et al.*, 1975; Sanchez *et al.*, 1995). Prevosti's distance between varieties g and g' using band information on M dominant markers is, $\delta_{g;g'}^{Prevosti} = \frac{1}{M}\sum_{m=1}^{M}\left|p_{mg} - p_{mg'}\right|$, which represents the average absolute band frequency difference between varieties g and g'. An estimator is created by replacing the parameters p_{mg} and $p_{mg'}$ in the expression for $\delta_{g;g'}$by their sample estimates $\hat{p}_{mg}$ and $\hat{p}_{mg'}$. The quadratic counterpart of Prevosti's distance coefficient is Rogers' distance, which is merely a Euclidean distance derived from band or allele frequencies, that is averaged over loci. For band frequencies Rogers' distance between varieties g and g' is $\delta_{g;g'}^{Rogers} = \sqrt{\frac{1}{M}\sum_{m=1}^{M}(p_{mg} - p_{mg'})^2}$, and again an estimator is obtained by replacing the population frequencies by their sample estimates. The estimator is biassed, but bias correction is straightforward (Ghérardi *et al.*, 1998). For allelic data from codominant markers, summation must take place over alleles and loci. Bias correction is then more elaborate (Lombard *et al*., 2001).

The null hypothesis for distinctness is that there is no difference between varieties g and g' with regard to the band (allele) frequencies. The sampling distribution of distances under the null hypothesis can be obtained by permutation (Ghérardi *et al.*, 1998). Within one permutation individual plants are allocated at random to varieties g and g'. Fingerprints, vectors of band presence/absence over the marker loci, are kept intact. Then band frequencies are calculated, followed by a distance estimate. To test equality of the band frequencies between varieties g and g', the original distance estimate for the non-permuted data is compared to the quantiles of the null distribution. When the original estimate is among the proportion α of most extreme values (largest distances) the null hypothesis of equal band frequencies is rejected and the varieties could be called distinct.

As an alternative to a permutation test for distinctness, we could try to construct a one-sided bootstrap confidence interval with a lower bound. It would be realistic to work with a minimum distance requirement, the null hypothesis is then $H_0: \delta_{g;g'} \leq \theta_D$, i.e., the real distance is smaller than a specified threshold for distinctness. The alternative hypothesis becomes $H_a: \delta_{g;g'} > \theta_D$, i.e., the real distance is larger. The threshold value, θ_D, is compared with the α quantile of the bootstrap distribution, and if θ_D is below the α quantile, the null hypothesis of equality is rejected and the varieties might be called distinct.

A test for genetic conformity can be constructed as the mirror image of the test for distinctness. For conformity the null hypothesis is $\delta_{g;g'} \geq \theta_{ED}$, i.e., the varieties are different, whereas the alternative hypothesis is $\delta_{g;g'} < \theta_{ED}$. A onesided bootstrap confidence interval is constructed with the upper bound given by the 1-α quantile of the bootstrap distribution. If the threshold, θ_{ED}, is larger than the 1-α quantile than the null hypothesis of difference is rejected and the varieties are concluded to be conform. The threshold θ_{ED} is again a matter of discussion. The conformity threshold, θ_{ED}, need not coincide with the distinctness threshold, θ_D.

3.3. Distinctness on the basis of distance information for individual pairs of plants

In subsections 3.1 and 3.2 procedures were described for distinctness and conformity using band or allele frequencies of varieties. In this subsection we present methods for analysing the difference and conformity of genetically heterogeneous varieties when distance estimates are available for each pair of individual plants. These distance estimates can be of any kind. For dominant marker systems producing bands, distances between individual plants might be calculated as 1 – Jaccard similarity, where the Jaccard similarity between two plants i and i' from variety g and g' is $s^{Jaccard}_{gi;g'i'} = \frac{n_{11}}{n_{11} + n_{10} + n_{01}}$, with n_{11} the number of bands present in both plant i of variety g and plant i'of variety g', n_{10} the bands present in only plant i of variety g, n_{01} the bands present only in plant i' of variety g'. Closely related to the Jaccard similarity are Nei and Li's similarity (Nei and Li, 1979), which is better known as Dice similarity in ecological literature, $s^{Dice}_{gi;g'i'} = \frac{2n_{11}}{2n_{11} + n_{10} + n_{01}}$, giving more weight to positive matches, and the simple matching coefficient, $s^{SM}_{gi;g'i'} = \frac{n_{11} + n_{00}}{n_{11} + n_{10} + n_{01} + n_{00}}$, that also counts negative matches, n_{00}. In words, Jaccard gives co-occurences of bands as fraction of the total number of bands in both varieties, Dice gives co-occurences as fraction of the arithmetic mean of occurences in both varieties, and simple matching gives co-occurences plus co-absences as fraction of the total number of plants over both varieties. Baril *et al*. (1997) pointed out that an argument in favour of Jaccard's similarity is that it does not depend on the individuals present in the samples of the varieties to be compared. In contrast, the simple matching coefficient does depend on the sample studied, but its use is more suitable for genetically close genotypes. For overviews of similarity measures, see Gordon (1981) and Digby and Kempton (1987). Similarity coefficients can be converted to distance measures by taking the complement, distance = 1 – similarity, or the square root of the complement, distance = $\sqrt{1 - \text{similarity}}$, where the square root variant is

Proc. Int. Symp. on Molecular Markers
Eds. Doré, Dosba & Baril
Acta Hort. 546, ISHS 2001

preferred as it confers the obtained distances the metric and Euclidean property, which is helpful when the distances are to be displayed graphically in a lower dimensional space as in principal coordinates analysis (see Digby and Kempton, 1987).

For allelic data, Rogers' or Prevosti's distance could be calculated between individual plants, where the frequencies p are replaced by presence/absence (1/0) indicators for individual alleles. For the methods in this subsection, distance measures do not necessarily have to be restricted to genetic band or allele frequency information. Mixed distance measures consisting of weighted sums of both genetic and phenotypic distances are completely feasible.

3.3.1. Analysis of distance

The analysis of distance (AOD) is based on the identity $\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n}\sum_{i<j}^{n}(x_i-x_j)^2$, that connects the corrected sum of squares for n observations with the sum over all pairwise distances between these observations. In genetics this principle was popularized by Excoffier *et al.* (1992) for comparing populations of genetically heterogeneous populations under the name of Analysis of Molecular Variance (AMOVA). In statistics the identity is used as the basis for a generalization of the Multivariate Analysis of Variance (MANOVA), such as to make possible MANOVA on sets of response variables that include both quantitative and qualitative response variables (Gower and Krzanowski, 1999).

We describe AOD for the general situation in which G groups are compared. By taking G=2 appropriate procedures for comparing two varieties originate. Let x_{gi} be the value on trait x of the i^{th} individual ($i=1...n_g$) in the g^{th} group (g=1...G), with $\bar{\bar{x}}$ the mean over all individuals and groups, and $\bar{x}_g$ the mean for group g. For a one-way ANOVA on x, the total variation $T=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gi}-\bar{\bar{x}})^2$ will be partitioned in a part due to differences between groups, $B=\sum_{g=1}^{G}n_g(\bar{x}_g-\bar{\bar{x}})^2$, and a residual, $W=\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{gi}-\bar{x}_g)^2$, representing variation within groups, or, T = B + W. When M variables, $x_1...x_M$ are considered simultaneously the latter decomposition can be summed over the variables, giving $T=\sum_{m=1}^{M}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{mgi}-\bar{\bar{x}}_m)^2$, $B=\sum_{m=1}^{M}\sum_{g=1}^{G}n_g(\bar{x}_{mg}-\bar{\bar{x}}_m)^2$, while $W=\sum_{m=1}^{M}\sum_{g=1}^{G}\sum_{i=1}^{n_g}(x_{mgi}-\bar{x}_{mg})^2$. A similar breakdown of the total variation in between and within variation can be obtained on the basis of squared Euclidean distances between individuals. Let the squared distance between an individual i in group g and an individual i' in group g' be $d^2_{gi;g'i'}=\sum_{m=1}^{M}(x_{mgi}-x_{mg'i'})^2$, the squared distance is the sum of the squared differences between the two individuals over all variables. The total variation, T, can be shown to be equal to the sum of all squared pairwise distances between individuals (over all groups) divided by the total number of individuals (over all groups). The within groups variations, W, is the sum over groups of the sum of squared pairwise distances within a group divided by the group size. The between group variation, B, can then be found by subtraction, B = T – W.

The partitioning of Euclidean distances over variables by AOD is equivalent to the partitioning of variation by ANOVA summed over variables. However, the key issue is that the partitioning according to AOD can be generalized to any kind of distance matrix, be it Euclidean or not, whether the traits are morphological, molecular, or a mixture, and whether the traits are quantitative, qualitative, or a mixture.

After the decomposition by AOD there still remains the problem of a testing procedure to assess whether, in case of distinctness, there are differences between the populations or not. For a test statistic, F, some options exist, like B/T, B/W, or $\frac{B/(G-1)}{W/(n-G)}$ (with n = total number of individuals). For distinctness testing G = 2, and $F = \frac{B}{W/n-2}$. The null distribution for the test statistics can once again be obtained by permutation of variety membership of individual plants. When the original test-statistic is among the proportion α most extreme values, the null hypothesis of equality of varieties is rejected at a test level α. For establishing a strict critical value it may be good to fix a level α critical value on the basis of the highest 1-α quantile found in the permutations for a complete set of reference and candidate varieties. This implies that a critical value could be defined once, where agreement on the appropriate set of varieties is required, after which new candidates could be compared on the basis of earlier set critical values.

3.3.2. Varland's multi-response test

Another test for a difference between genetically heterogeneous varieties, which uses the matrix of pairwise distances between the individuals of different varieties is Varland's multi-response permutation test (see Manly 1997, pages 263-264). The idea is to correlate a distance (or similarity) matrix derived from morphological and/or marker information with a distance matrix derived from variety membership, i.e., individuals belonging to the same variety have distance 0, otherwise the distance is 1. Varland's test is then essentially a Mantel test, a permutation test, for the correlation (or covariance) between the off-diagonal elements of the two distance matrices. Effectively the variety membership of individual plants is permuted, and after each permutation a new distance matrix for variety membership is calculated. Subsequently, the correlation is calculated between membership's distance and traits' distance. Finally, the original correlation is compared to the quantiles of the null distribution obtained from permutation, and based on the extremity of the original correlation the null hypothesis of equal varieties is rejected or accepted.

A slight modification on this test can be to fit a logistic regression with the membership distances as responses, which can take only the values 0 or 1, and the trait distances as predictors. The test statistic can be the t-value for the slope, whose significance can be assessed by a permutation test.

3.4. Distinctness and conformity on the basis of distance information for individual pairs of plants, without resampling

Instead of using permutation or bootstrap procedures for testing the difference between two varieties, one can also define an explicit model for the pairwise distance or similarities between individuals in different varieties. Curnow (1998) proposed a model inspired by models for the phenotypic response of offspring produced from factorial mating designs. The assumption is made that these distances are observed without error. For the distance between an individual i in variety g and an individual i' in variety g',

$d_{gi;\,g'i'} = \mu + Z_i^g + Z_{i'}^{g'} + Z_{ii'}^{gg'}$. Under the assumption that Z_i^g, $Z_{i'}^{g'}$, and $Z_{ii'}^{gg'}$ are independent random effects whose variances are $\sigma_g^2, \sigma_{g'}^2, \text{and} \sigma_{g;g'}^2$, the variance of the average distance between the varieties g and g' is $\text{var}(\bar{d}_{g;g'}) = \frac{n_g \sigma_{g'}^2 + n_{g'} \sigma_g^2 + \sigma_{g;g'}^2}{n_g n_{g'}}$. The variances $\sigma_g^2, \sigma_{g'}^2, \text{and} \sigma_{g;g'}^2$ can be estimated by equating expected and observed mean squares of an ANOVA on the two-way table of pairwise distances with rows corresponding to individuals from variety g and columns corresponding to variety g'. Fitting an additive two-way model to the pairwise distances provides:

- the mean square between individuals from variety g (the row factor), with expected mean square, $\sigma_{g;g'}^2 + n_{g'} \sigma_g^2$
- the mean square between individuals from variety g' (the column factor), with expected mean square, $\sigma_{g;g'}^2 + n_g \sigma_{g'}^2$
- a residual, $\sigma_{g;g'}^2$

The estimate for $\text{var}(\bar{d}_{g;g'})$ can be used to create confidence intervals for $\bar{d}_{g;g'}$. Hypotheses concerning distinctness and conformity can then be formulated along the ways indicated in subsection 3.2.

4. Statistical procedures for assessing distinctness and conformity for genetically homogeneous varieties

For assessing distinctness between pairs of homogeneous varieties on the basis of molecular markers, the paradoxical situation occurs that a one locus difference could be enough when no explicit minimum distance requirement is formulated. The situation is comparable to that for assessing distinctness in many ornamentals where a different flower colour is often enough for the granting of plant breeders' rights. When there is no statistical variation, i.e., all individuals of a variety are genetically the same, there is no need for the implementation of statistical procedures for distinctness. It is without controversy that markers may serve well for *identification* of especially homogeneous varieties, the markers then provide a kind of barcode. For *registration*, however, markers should exhibit a direct link with functional genes that code for traits that are already in use for distinctness (Camlin, 2001). As an example we can think of a marker for the major gene resistance against Bremia in lettuce. The marker then merely replaces the already accepted phenotypic trait. Such an implementation of markers can be desirable for cost reduction purposes.

An approach using molecular markers for distinctness between genetically homogeneous varieties without explicitly defining a minimum distance will not be very productive, as it will evidently lead to a severe erosion of plant variety protection. Work by Law *et al.* (1998) acknowledges this problem and proposes that discrimination on the basis of markers should be calibrated on discrimination by phenotypic traits. AFLP fingerprint profiles for wheat varieties were compared at increasing stringency of a distinctness criterion, where the criterion consisted in the number of bands at which two varieties should differ for distinctness. This seems a sensible approach for finding an explicit minimum distance attached to the use of molecular markers. The proviso to be made is that such an approach will have to take care of discrimination by markers being comparable to discrimination by phenotypic characters, not only with respect to the proportion of variety pairs being distinguished, but also with respect to which pairs are being distinguished and which not. To comply with the latter condition, markers linked to the UPOV prescribed phenotypic characters would be the best candidates. When distinctness is defined by two varieties exhibiting more than a threshold proportion of

difference at either a fixed or random number of marker loci, conformity could similarly be defined by having less than a threshold proportion of differences. Again, the most natural way to find such a threshold would be empirical by investigating the proportion of differences for variety pairs of known relationship.

A more statistically inspired approach towards distinctness and conformity testing for homogeneous varieties using band data is given by Lombard *et al.* (2001). They make the assumption that the number of co-occurences of bands, n_{11}, is a binomial variable with sample size $(2n_{11} + n_{10} + n_{01})/2$ and success rate $2n_{11}/(2n_{11} + n_{10} + n_{01})$. From this they derive the distribution of Nei and Li's (Dice) similarity. With this distribution tests for distinctness and conformity can be constructed immediately (see section 3.2 for the principle), without the necessity of resampling.

For allelic data Dillmann *et al.* (1997b) proposed a procedure for homogeneous varieties that are also homozygous. For such varieties, Rogers' distance, $\delta_{g;g'}^{Rogers}$, measures the proportion of loci at which two varieties differ. An upper bound for the standard error of this distance would be given by $\sqrt{\frac{\delta_{g;g'}^{Rogers}(1-\delta_{g;g'}^{Rogers})}{M}}$, when all M markers are assumed to be independent. A better approximation to the standard error of the distance will be obtained by taking into account the distances between marker loci. An expression for such a standard error is given by Dillmann *et al.* (1997b). Still, it appears that the approach needs the rather restrictive assumption of the probability of a band being the same for all marker loci.

For finding standard errors of distance measures for pairs of homogeneous varieties, bootstrap procedures resampling loci within variety profiles emerge as an attractive option. However, bootstrap procedures require independence of the sampling units and it will be clear that marker loci will not be independent due to linkage disequilibrium and linkage (van Dongen, 1995). Bootstrap procedures may work when markers can be chosen from different chromosomes, or at large distance within chromosomes.

As a generally useful alternative to the above described procedures for homogeneous varieties, the forensic testing approach developed by especially Weir and co-workers can be used (Weir, 1996; Evett and Weir, 1998). An application to plant variety protection is given by Ibañez (2001). In forensic testing the likelihood is estimated of a DNA profile found at the place of a crime, the perpetrator profile, being the same as the profile of a suspect. In its simplest form, assuming perpetrator and suspect have independent profiles when they are different people, this likelihood is given by the reciprocal of the probability for the perpetrator profile. This probability can easily be calculated when Hardy-Weinberg and linkage equilibrium are the case. The probability of a profile can then be calculated as the product of the probabilities for specific genotypes at individual loci, which are p_{ma}^2 for the homozygote with allele a at marker m, and $2p_{ma}p_{ma'}$ for the heterozygote with alleles a and a' at locus m.

The principle of calculating likelihoods for particular profiles is very suitable for answering questions of genetic conformity for crops in which new varieties can be developed by selection of mutants and sports. It is to be expected that when a new variety is just a mutant of another variety, the profiles of both will very probably be identical unless markers were targeted at identifying specific mutations. The question that then arises is: what is the probability of finding an identical profile for a 'new' variety and an existing, protected variety? This probability and the corresponding likelihood can be calculated from knowledge of allele and/or genotype distributions at individual loci in a reference population of genotypes, together with knowledge on distance/ independence between loci. The definition of the reference population of genotypes determines the distributions of alleles and genotypes. The likelihood principle is also useful for assessing conformity between varieties that are not expected to have identical profiles. In that case

the new variety and the older, protected variety could be allowed to differ at a certain proportion of either or both a set of fixed and random loci. The calculation of the likelihood will not become more complicated by this variation. In addition, the likelihood principle could be extended even further to heterogeneous varieties, although it is questionable whether a likelihood approach for these varieties would be more powerful than the procedures described in section 3.

5. Discussion

5.1. Number of markers

For successful implementation of marker technology in distinctness and conformity testing, it is important to know the number of markers needed for reliable assessment. The quality of a marker application can be quantified by the size of the average standard error, or the length of the average confidence interval, for a distance measure in relation to the average distance to be expected (assuming unbiassed estimators or estimators with a constant bias). Efficiency of a marker system becomes higher when the ratio expected distance / standard error increases. Leaving aside laboratory errors, the standard error for a distance estimator will depend on the number of markers, the number of alleles per marker, the distributions of the alleles within loci, the dispersion of the markers over the genome, and the real distance. For genetically heterogeneous varieties, an additional factor is the number of plants per variety. Results on numbers of bi-allelic or multi-allelic loci, and numbers of plants necessary for acceptable standard errors are scarce. Derivations on standard errors all use the simplifying assumptions of independent marker loci and linkage disequilibrium, and by that avoid the problem of the dispersion of the markers over the genome. Furthermore, the loci are assumed to be 'exchangeable', which loosely translated means that the alleles within loci are supposed to follow the same distribution from one locus to the next. This condition excludes loci which are under selection. The theory thus refers mainly to selectively neutral loci in populations that exercise random mating. We summarize some results for genetically heterogeneous varieties.

An explicit expression for the standard error of a (squared) distance based on quadratic differences in allele frequencies that are standardized by dividing by average frequencies is given in Foulley and Hill (1999). The definition of the distance for one locus with A alleles is $\delta^2_{g;g'} = \sum_{a=1}^{A} \frac{(p_{ag} - p_{ag'})^2}{(p_{ag} + p_{ag'})/2}$, for M loci the average of this expression over loci is taken. For M independent loci with A alleles, N plants per variety and real distance $\delta^2_{g;g'}$ between the varieties, the standard error for the estimator is $SE(\hat{\delta}^2_{g;g'}) = \sqrt{\frac{2}{M(A-1)}}\left[\delta^2_{g;g'} + \frac{1}{N}\right]$. For example, choosing M = 20, which is a reasonable number of independent loci on a genome of 1000 cM, A = 2, meaning loci are bi-allelic, N = 20, which is an acceptable number of plants per variety for profiling, and $\delta^2_{g;g'} = 0.0625$, an analogon with the expected genetic relatedness between recurrent parent and back cross offspring after 3 generations of back crossing, the standard error will be 0.0356. This means that a confidence interval will span about +/- 0.07, quite a lot for a real distance of 0.0625. Equivalent to 20 bi-allelic loci would be 5 multi-allelic loci with 5 alleles. Increasing the number of plants to 100 would reduce the standard error to 0.0229. If it would be possible to increase the number of independent bi-allelic loci to 80, or more realistically, to increase the number of independent multi-allelic loci with 5

alleles to 20, the standard error would become 0.0179. Increasing the number of loci and/or the number of alleles per locus is thus more effective than increasing the number of plants per variety.

Related work was presented by Ghérardi *et al.* (1998) for a bias corrected form of Rogers' distance, that is also based on quadratic differences in allele frequencies, $(p_{ag} - p_{ag'})^2$, but without the standardization by average frequency. Interpolating their Fig. 2, a standard error for a squared distance of (roughly) 0.06 would be approximately 0.035, when about 20 bi-allelic markers are used in combination with about 20 plants per variety. Foulley and Hill's formula gives the same magnitude of standard error for that situation.

Other interesting work was done by Lynch and Ritland (1999) on estimating genetic relatedness for pairs of *individuals* (not varieties), where genetic relatedness is defined as twice the coefficient of coancestry (the probability that a random gene from one individual is identical by descent with a random gene from another individual). Their results provide additional insights on what to expect in terms of errors. For their best estimator for genetic relatedness and under the assumption of the most favourable distribution of alleles within loci (uniform) plus again the assumption of independent loci, they give for *distant* relations a minimum attainable error of $\sqrt{\frac{1}{M(A-1)}}$, where M is the number of markers and A is the number of alleles. For parent-offspring and multi-allelic loci, the standard error can be reduced up to 50% in comparison to these quantities for distantly related individuals. Note that the effects of the number of marker loci and the number of alleles per locus are accounted for in the same way as in the formula of Foulley and Hill (1999). As an example, for 20 multi-allelic loci with 5 alleles, the standard error for a parent-offspring relation would be about 0.055. For recurrent-parent back cross offspring the standard error will be smaller, although it is difficult to tell by how much.

What can be concluded from these studies is that for distances that might occur in disputes on essential derivation, standard errors around 0.02, and confidence intervals of 0.08, constitute a minimum level of variation. When a conformity threshold is put at 0.90 (θ_{ED} = 0.10), pairs of varieties with real conformities of 0.86 can still surpass this threshold by coincidence.

The standard errors above were all derived for the case of heterogeneous varieties. For homogeneous varieties that are in addition homozygous, and under the assumption of exchangeability of loci (independently and identically distributed), the standard error for Rogers' distance is given by an application of the binomial law (see section 4). For a distance of 0.0625 and 20 independent marker loci the standard error will then amount to 0.0541. Dillmann *et al.* (1997b) developed a best linear unbiassed estimation (BLUE) procedure that takes into account dependency between loci. The corresponding estimator for the standard error can in principle be used for calculating numbers of required markers for given constellations of markers over the genome. Lombard *et al.* (2000) extended this line of work. For rapeseed varieties with on average one marker per 18 cM, with a BLUE procedure, modelling the dependencies between marker loci, a gain in precision of 23% was achieved. When these markers would have been distributed equidistantly over the genome, the gain would have been 40%.

5.2. Random or fixed markers

A frequently asked question with respect to measuring distinctness and conformity is what to choose, a random set of markers, whose positions are unknown, or a fixed set of markers, whose positions are, to a certain extent, known? The basic issue is to take care to sample the genome in a 'representative' way. When markers appear in clusters, it

Proc. Int. Symp. on Molecular Markers
Eds. Doré, Dosba & Baril
Acta Hort. 546, ISHS 2001

seems wise to downweight individual markers in such clusters. Some distance measures take automatically care of such dependencies (see Nei, 1987, p. 212-214), and this might be an argument in favour of such measures. The price to be paid for this automatic weighting is usually that of more complicated distributions for test statistics. When the position of markers is more or less known, the optimal sampling scheme will be a scheme in which for a given number of loci, loci are equidistantly sampled within the genome. A good starting configuration would be to have one marker locus at each chromosome arm, as this may provide a maximum set of independent loci.

When no markers with known positions are available, and a random set of markers must be used, it is sensible to optimize the composition of the marker set. We want to estimate distances as precisely as possible using as few resources as necessary. Various methods have been presented for optimizing the information content of a set of markers. Individual loci are often ranked on information content by their average heterozygosity value or gene diversity $1-\sum_{a=1}^{A} p_a^2$ (Weir, 1996). However, heterozygosity values are useful for choosing one marker among a group of markers, but they do not guide in the construction of a marker set aiming at minimizing standard errors for distances. An acceptable solution to the problem of marker set construction is to perform a multiple regression of a variety membership indicator (for example, 1 for plants from variety g and 0 for plants from variety g'), on a predictor set consisting of band or allele indicators (1 for presence, 0 for absence). A subset selection procedure like stepwise forward can be used to determine a minimum set of markers with complementary information. In the search for a best subset of markers, allele indicators belonging to the same locus should be entered or removed jointly. The reason for regression performing well is that for the two groups situation regression with an indicator for group membership is equivalent to discriminant analysis, where discriminant analysis is the standard technique for constructing and identifying discriminatory variables. Preselection of markers by stepwise regression led to more powerful tests for distinctness for the perennial ryegrass varieties used in Roldán-Ruiz *et al.* (2001).

For homogeneous varieties a possibility is to perform a correspondence or principal components analysis (Digby and Kempton, 1987) on the marker data combining all reference and candidate varieties. Based on a biplot of the markers, a number of markers can be chosen that act complementary over the whole of the set of varieties.

5.3. Relations between marker, morphological and pedigree information

The relationship between marker information and morphological information has been the subject of investigation in a number of studies over the last years. An unequivocal association would raise the acceptance level of markers as additional characters for distinctness. When marker assessments could replace phenotypic observations, cost reductions are within reach. Theoretical results of Burstin and Charcosset (1997) suggest that the relationship between morphological and marker information is most likely triangular, close genetic relationships correspond necessarily with close morphological relationships, whereas distant genetic relationships can correspond with both close and distant morphological relationships. Observations on marker – trait distances both support (Dillmann *et al.*, 1997a; Burstin and Charcosset, 1997) and contradict (Roldán-Ruiz *et al.*, 2001) the triangular view. In general, for close marker based relationships the association with morphology seems reasonable. Nuel *et al.* (2001) showed for maize how morphological distances can be predicted from marker based distances, and that a procedure using predicted morphological distances offers a viable alternative to classical distinctness testing (see Nuel *et al.*, 2000, for statistical

details). Because of the good correspondence between close marker distances and close morphological distances, markers could also play an important role in the management of reference collections. When marker information correctly reflects morphological information it may be possible to reduce the number of reference varieties actually tested in the field, as then for each candidate variety its most similar reference varieties could be selected beforehand in conformance with the molecular information (Dillmann and Guérin, 1998; Law *et al.*, 1999).

Where a tight association between marker based distances and morphological distances is important for distinctness purposes, for conformity the relation between marker based distances and pedigree distances, or coancestry, is important. In the early days of the essential derivation concept it was believed that genetic similarities as calculated from marker information could straightforwardly be interpreted as estimators of pedigree relations. The implementation of a conformity threshold would then contain little more than a discussion about whether a third or a fourth back cross should be declared essentially derived. Subsequently, any genetic similarity with good sampling properties in the sense of having small variance would suffice for establishing conformity beyond the predetermined threshold. As indicated by Lynch (1988), genetic similarity arises due to alleles being identical in state and identical by descent. Pedigree relations are expressed in terms of identity by descent only. The problem to be solved now is to separate out from a genetic similarity estimate the part due to identity by descent (coancestry) and the part due to identity in state that is not due to coancestry (selection, drift, mutation, etc..). One solution was put forward by Lynch (1988). Assuming random mating and no inbreeding, the relation between genetic similarity and genetic relatedness (expected fraction of genes in variety g that is identical by descent with those in variety g') is $S_{g;\,g'} = r_{g;\,g'} + (1 - r_{g;\,g'})S_0$, or, the expected similarity between g and g', $S_{g;\,g'}$, is the sum of a fraction genes identical by descent, $r_{g;\,g'}$, and a complementary fraction, $(1 - r_{g;\,g'})$, that is not identical by descent, but for various reasons is identical in state, with a probability given by S_0, the expected similarity between non-relatives. One major problem in finding estimates for genetic relatedness (pedigree) is that good estimates for S_0 are difficult to obtain, as we usually do not know which varieties are unrelated to the pair we are investigating. The distribution of genetic similarity measures is strongly dependent on the allele frequency distributions within marker loci over the reference population. For estimating pedigree relationships from genetic similarities, first of all the reference population of varieties should be described, then information is necessary about the marker allele frequency distributions, the mating system, and the type and intensity of selection exercised. The curtailment of the reference population will by no means be simple as it should comprehend all varieties of a crop for which comparable cases of essential derivation could occur. Besides the definition of the reference population, the effects of selection create serious complications in the estimation of pedigree relationships from genetic similarities. These problems were found more serious for autogamous crops like barley and wheat than for an allogamous crop like maize by Bohn *et al.* (1999). As an explanation they propose that for barley and wheat segregants are selected for phenotypic similarity with the preferred parent, whereas for maize selection is performed on combining ability without paying much attention to the phenotypic similarity with either of both parents.

Although methods have been proposed that might unravel identity in state from identity by descent in the face of selection (Bernardo *et al.*, 1996), it still seems that approaching essential derivation via genetic relatedness creates prohibitive complications. Conceptually the most transparent solution to assessing genetic conformity for essential derivation purposes, is by the construction of distributions for genetic similarity or distance estimators for known relationships. These relationships should be agreed upon examples of 'initial variety – essentially derived variety', and some control relations (distant, parent-offspring, F1, BC1). The procedure is to generate samples of variety pairs with a defined relation, and then just calculate distances for this relationship. The set of distances corresponding to a particular relation determines the empirical distribution for

Proc. Int. Symp. on Molecular Markers
Eds. Doré, Dosba & Baril
Acta Hort. 546, ISHS 2001

the relation given the marker system and protocol. When in future doubts might arise on the non-derivation of certain varieties, the distances of these new varieties to potential initial varieties should be measured, and the obtained distances could be compared with the empirical distance distributions for a number of known relationships. The approach outlined in these last lines is momentarily under study in the EU-MMEDV project and the ASSINSEL essential derivation study group for lettuce.

Acknowledgement

Work of the first author was funded by the EU project 'Molecular and other markers for establishing essential derivation (EDV) in crop plants' (QLRT-1999-PL1499). Work of the second author results from projects funded by the French Ministry of Agiculture.

References

Baril C.P., Verhaegen D., Vigneron P., Bouvet J.-M., and Kremer A., 1997. Structure of specific combining ability between two species of *Eucalyptus*. I. RAPD data. Theor. appl. Genet., 94: 796-803.

Bernardo R., Murigneux A., and Karaman Z., 1996. Marker-based estimates of identity by descent and alikeness in state among maize inbreds. Theor. appl. Genet., 93: 262-267.

Bohn M., Utz H.F., and Melchinger A.E., 1999. Genetic similarities among winter wheat cultivars determined on the basis of RFLPs, AFLPs, and SSRs and their use for predicting progeny variance. Crop Sci., 39: 228-237.

Burstin J., and Charcosset A., 1997. Relationships between phenotypic and marker distance: theoretical and experimental investigations. Heredity, 78: 477-483.

Camlin M. S., 2001. Possible future roles for molecular techniques in the identification and registration of new plant cultivars. Acta Hort 546.(this issue).

Curnow R.N., 1998. Estimating genetic similarities within and between populations. J. Agric. Biol. Env. Statist., 3: 347-358.

Digby P.G.N., and Kempton R.A., 1987. Multivariate analysis of ecological communities. Chapman and Hall Ltd., London.

Dillmann C., Bar-Hen A., Guérin D., Charcosset A., and Murigneux A., 1997a. Comparison of RFLP and morphological distances between maize *Zea mays* L. inbred lines. Consequences for germplasm protection purposes. Theor. appl. Genet., 95: 92-102.

Dillmann C., Charcosset A., Goffinet B., Smith J.S.C., and Dattée Y., 1997b. Best linear unbiased estimator of the molecular genetic distance between inbred lines. Advances in Biometrical Genetics. Proceedings of the Tenth Meeting of the EUCARPIA Section Biometrics in Plant Breeding, Poznan , 14-16 May 1997, P. Krajewsky and Z. Kaczmarek (eds.), pp 105-110.

Dillmann C., and Guérin D., 1998. Comparison between maize inbred lines: genetic distances in the expert's eye. Agronomie, 18: 659-667.

Efron B.J., and Tibshirani R.J., 1993. An introduction to the bootstrap. Chapman and Hall, London.

Evett I.W., and Weir B.S., 1998. Interpreting DNA evidence, Statistical genetics for forensic scientists. Sinauer, Sunderland.

Excoffier L., Smouse P.E., and Quattro J.M., 1992. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. Genetics, 131: 479-491.

Foulley J.-L., and Hill W.G., 1999. On the precision of estimation of genetic distance. Genet. Sel. Evol., 31: 457-464.

Ghérardi M., Mangin B., Goffinet B., Bonnet D., and Huguet T., 1998. A method to measure genetic distance between allogamous populations of alfalfa (*Medicago sativa*) using RAPD molecular markers. Theor. appl. Genet., 96: 406-412.

Gilliland T.J., Coll R., Calsyn E., De Loose M., Van Eijk M.J.T., and Roldán-Ruiz I., 2000. Estimating genetic conformity between related ryegrass (*Lolium*) varieties, I. Morphology and Biochemical Characterisation. Mol Breed. (in press).
Gordon A.D., 1981. Classification. Chapman and Hall, London.
Gower J.C., and Krzanowski W.J., 1999. Analysis of distance for structured multivariate data and extensions to multivariate analysis of variance. Appl. Statist., 48: 505-519.
Ibañez J., 2001. Mathematical analysis of RAPD data to establish reliability of varietal assignment in vegetatively propagated species. Acta Hort 546,(this issue).
Law J.R., Donini P, Koebner R.M.D., Reeves J.C.. and Cooke R.J., 1998. DNA profiling and plant variety registration. 3: The statistical assessment of distinctness in wheat using amplified fragment length polymorphisms. Euphytica 102: 335-342.
Law, J.R., Cooke R.J., Reeves J.C., Donini P., and Smith, J.S.C., 1999. Most similar variety comparisons as a grouping tool. Plant Varieties and Seeds, 12: 181-190.
Lombard V., Baril C.P., and Zhang D., 2000. Improvement of the precision of genetic estimates between rapeseed cultivars based on AFLP by using information from their position on a consensus linkage map. Plant and animal genome congress, San Diego, USA, 9-13 january.
Lombard V., Dubreuil P., Dillmann C., and Baril C.P., 2001. Genetic distance estimators based on molecular data for plant registration and protection: a review. Acta Hort 546, (this issue).
Lynch, M. 1988. Estimation of relatedness by DNA fingerprinting. Mol. Biol. Evol., 5: 584-599.
Lynch M., and Ritland K., 1999. Estimation of pairwise relatedness with molecular markers. Genetics, 152: 1753-1766.
Manly B.F.J., 1997. Randomization, bootstrap and Monte Carlo methods in biology, 2nd Edition, Chapman and Hall, London.
Nei M., 1987. Molecular evolutionary genetics. Columbia University Press, New York.
Nei M., and Li W.H., 1979. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Science of the United States of America, 76, 5269-73.
Nuel G., Robin S., and Baril C.P., 2000. Predicting distances using a linear model: the case of varietal distinctness. Journal of Applied Statistics (in press).
Nuel G., Baril C.P. and Robin S., 2001. Varietal distinctness assisted by molecular markers: a methodological approach. Acta Hort 546, (this issue).
Prevosti A., Ocaña J., and Alonso G., 1975. Distances between populations of Drosophila subobscura based on chromosome arrangement frequencies. Theor. appl. Genet., 45: 231-241.
Roldán-Ruiz I., Calsyn E., Gilliland T.J., Coll R., Van Eijk M.J.T., and De Loose M., 2000. Estimating Genetic Conformity Between Related Ryegrass (*Lolium*) varieties, II. AFLP Characterisation. Mol. Breed. (in press).
Roldán-Ruiz I., van Eeuwijk F.A., Gilliland T.J., Dubreuil P., Dillmann C., Lallemand J., De Loose M., and Baril C.P., 2001. A comparative study of molecular and morphological methods of describing relationships between perennial ryegrass (*Lolium perenne* L.) varieties. Theor. appl. Genet., (in press).
Sanchez A., Ocaña J., and Utzet F., 1995. Sampling theory, estimation, and significance testing for Prevosti's estimate of genetic distance. Biometrics, 51: 1216-1235.
UPOV, 1991. International convention for the protection of new varieties of plants, Geneva, Publication No. 221 (E), March 19, 1991.
UPOV, 1997a. Use of COYD and COYU. TWC/15/6.
UPOV, 1997b. User notes for combined-over-years distinctness and uniformity procedures. TWC/15/7.
Van Dongen S., 1995. How should we bootstrap allozyme data? Heredity, 74: 445-447.
Van Dongen S., and Backeljau T., 1995. One- and twosample tests for single-locus inbreeding coefficients using the bootstrap. Heredity, 74: 129-135.
Weir B.S., 1996. Genetic data analysis II. Sinauer, Sunderland.