J. G. Tivang · J. Nienhuis · O. S. Smith

Estimation of sampling variance of molecular marker data using the bootstrap procedure

Abstract Knowledge of genetic relationships among genotypes is useful in a plant breeding program because it permits the organization of germplasm and provides for more efficient sampling. The genetic distance (GD) among genotypes can be estimated using random restriction fragment length polymorphisms (RFLPs) as molecular markers. Knowledge of the sampling variance associated with RFLP markers is needed to determine how many markers are required for a given level of precision in the estimate of GD. The sampling variance for GD among all pairs of 37 maize (Z. mays L.) inbred lines was estimated from 1,202 RFLPs. The 1,202 polymorphisms were generated from 251 enzyme-probe combinations (EPC). The sampling variance was used to determine how large a sample of RFLPs was required to provide a given level of precision. The coefficient of variation (CV) associated with GD has a nearly linear relationship between its expected standard deviation and mean. The magnitude of the decrease in the mean CV for GD with increasing numbers of bands was dependent upon the sampling unit; e.g., individual polymorphic bands vs EPC, and the degree of relatedness among the inbreds compared. The rate of reduction in mean CV with increasing sample size was the same regardless of the restriction enzyme used, BamHI, EcoRI or HindIII, when the bootstrap sampling units were individual polymorphic bands. In contrast, although the rate of reduction (slopes) was the same, the intercepts of the mean CVs were different when EPCs were used as the bootstrap sampling unit. This difference was due to the higher number of bands per EPC in BamHI (4.94) compared with EcoRI (4.83) and HindIII (4.63).

Key words RFLP · Bootstrap · Sampling variance · Zea mays

Introduction

Knowledge of genetic relationships among genotypes is useful in a plant breeding program because it permits the organization of germplasm and provides for more efficient sampling. Molecular markers, including restriction fragment length polymorphisms (RFLPs) and random amplified polymorphic DNAs (RAPDs), have been used to estimate the genetic relationships among genotypes in numerous crop species, including maize (Lee et al. 1989; Godshalk et al. 1990; Smith et al. 1990; Messmer et al. 1992), Brassica species (Song et al. 1988; Slocum et al. 1990; Hu and Quiros 1991; Nienhuis et al. 1993), Cucumis melo (Neuhausen 1992), Solanum species (Debener et al. 1990), and Cucurbita (Wilson et al. 1992). The number of polymorphic molecular marker bands used to estimate genetic relationships varies widely from 61 (Nienhuis et al. 1993) to 1,205 (Smith et al. 1990).

Sampling variance in the estimation of genetic relationships occurs when discrepancies are detected between a random subset of molecular marker bands and all possible bands. Larger numbers of random polymorphic bands will provide an increasingly more precise estimate of genetic relationships and will reduce the variance caused by over- or under-sampling certain regions of the genome. Moreover, if marker bands could be chosen to be uniformly distributed over linkage groups, the sampling variance due to over- or under-sampling could be reduced. Nevertheless, to obtain larger numbers of random polymorphic bands is now expensive in terms of time and resources, and linkage information among polymorphic bands is not always available. Thus, it would be desirable to estimate genetic relationships using the smallest set of polymorphic bands which minimize sampling variance (Smith et al. 1990).
The bootstrap is a computer-intensive sampling method designed to empirically estimate variance where theoretical computations may be problematic (Felsenstein 1985; Efron and Tibshirani 1986, 1991; Krajewski and Dickerman 1990). The bootstrap procedure has been used by Felsenstein (1985) to estimate the sampling variance associated with phylogenetic analysis. Molecular markers usually generate large amounts of data, providing an excellent opportunity for bootstrap sampling both of whole data sets and within smaller partitions of the data set. Portions could be based on the degree of relatedness among the inbreds or on the restriction enzymes used in RFLP analysis.

The objectives of the present study were: (1) to determine the sampling variance using different size bootstrap samples from a large (1202 bands) RFLP data set; (2) to estimate the sampling variance for different sampling units (e.g., individual polymorphic bands or enzyme probe combinations); and (3) to compare the bootstrap sampling variance associated with the use of different restriction enzymes.

Materials and methods

Germplasm and RFLP data

Thirty-seven elite Zea mays inbred lines (Pioneer Hi-bred International, Johnston, Iowa) were used in this study. A detailed description of the inbreds is provided in previous publications (Smith and Smith 1989; Smith et al. 1990). The methods for DNA isolation, restriction endonuclease digestion, electrophoresis, Southern blotting, hybridization and autoradiography have been previously described (Smith et al. 1990). The restriction fragments were scored from the autoradiograms resulting in a matrix of 37 inbred lines and 1202 bands. The 1202 bands were generated from 251 mapped enzyme-probe combinations which were evenly distributed across the genome (Smith et al. 1990).

Terminology and definitions

The term enzyme-probe combination (EPC) is equivalent to "restriction enzyme probe combinations" (Smith et al. 1990), "clone-enzyme combinations" (Lee et al. 1989; Messmer et al. 1992), and "probe/enzyme combinations" (Debener et al. 1990), while band is equivalent to polymorphic/monomorphic band. The presence of a restriction enzyme probe combination was scored as "1", while absence of a fragment was scored as "0" (Lansman et al. 1981). The 1s and 0s are referred to as scores. The term "fragment frequency p(1)" refers to frequency of a type "1" score.

Statistical analyses of data

A computer program, written in Think C (Symantic, Cupertino Calif.), was developed to read in the data file, and to calculate the genetic distance (GD). Estimates of GD were calculated for all 666 pairs of inbreds according to the following equation which is the complement of the simple matching coefficient (Gower 1985):

\[
GD(i,j) = \frac{\sum N(i \neq j)}{(\sum N(i \neq j) + \sum N(i = j))},
\]

where GD is the measure of genetic distance between inbreds i and j, while \(\sum N(i \neq j)\) and \(\sum N(i = j)\) are the total number of scores discordant and concordant between inbreds i and j, respectively, over all N bands considered. A GD value of 0.0 and 1.0 indicates, respectively, no and maximal RFLP difference between two inbreds.

Bootstrap

To empirically estimate the sampling variance, a computer program was written to execute a bootstrap sampling procedure which is an adaptation of the procedures of Efron and Tibshirani (1986) and Felsenstein (1985). The program was designed to execute bootstrap sampling, using either individual bands or EPCs. Two hundred subsets, for a given number of N polymorphic bands, were sampled. The N bands were selected at random from the whole data set of 1202 polymorphic bands. Sampling with replacement generated a probability of 1/1202 for any one band to be selected. A second bootstrap sampling method was also used, where the sampling unit was the the EPC. In this case 200 subsets of M arbitrary EPCs were sampled. Random selection from the whole data set with replacement generated a probability of 1/251 for any one EPC to be selected at any one time. The GD was calculated between all (666) inbred pairs. The variability among 200 bootstrap samples for each pair of inbreds was standardized using the coefficient of variation (CV). Normalization of the variance to the CV was possible since the relationship between the variance and the mean for the GD estimator in the targeted range is nearly linear (Tivang 1992).

The mean CV of the 666 inbred pairs for a given sample size was plotted against the sample size. When the EPC was used as the sampling unit, the mean and standard deviation for the number of bands in each sample was calculated. The mean CV was plotted against the mean number of bands included in each sample of M EPCs. Linear regression calculations were obtained by natural logarithmic transformation of mean CV and bootstrap sample size.

Bernoulli random variable

The GD statistic is based on two outcomes for each comparison; either the scores agree (00, 11) or disagree (01, 10) for any two inbreds. This dual outcome can be modelled as a Bernoulli random variable. The probability of observing an agreement or disagreement is then a function of the fragment frequency.

Results

A total of 251 enzyme probe combinations (EPC) were scored for each of the three restriction enzymes BamHI, EcoRI, and HindIII, with a mean of 4.79 ± 0.06 (polymorphic) bands per EPC (Table 1). Among the 37 inbreds and the 1202 polymorphic bands the probability of finding a fragment present (1) and absent (0) was 0.249 and 0.751, respectively.

Comparison of sampling units

Under the assumption that RFLP data can be modelled as a Bernoulli random variable, the following statistics

<table>
<thead>
<tr>
<th>Restriction enzyme</th>
<th>Number of EPC (probes)</th>
<th>Number of bands</th>
<th>Mean number of bands per EPC</th>
</tr>
</thead>
<tbody>
<tr>
<td>BamHI</td>
<td>78</td>
<td>385</td>
<td>4.94 ± 2.21</td>
</tr>
<tr>
<td>EcoRI</td>
<td>82</td>
<td>396</td>
<td>4.63 ± 2.21</td>
</tr>
<tr>
<td>HindIII</td>
<td>91</td>
<td>421</td>
<td>4.63 ± 1.81</td>
</tr>
<tr>
<td>Total</td>
<td>251</td>
<td>1202</td>
<td>4.79 ± 2.06</td>
</tr>
</tbody>
</table>