Theor Appl Genet (2007) 114:421–428
DOI 10.1007/s00122-006-0433-9

ORIGINAL PAPER

Genetic distance sampling: a novel sampling method for obtaining core collections using genetic distances with an application to cultivated lettuce

J. Jansen · Th. van Hintum

Received: 3 January 2006 / Accepted: 12 October 2006 / Published online: 16 December 2006

Abstract This paper introduces a novel sampling method for obtaining core collections, entitled genetic distance sampling. The method incorporates information about distances between individual accessions into a random sampling procedure. A basic feature of the method is that automatically larger samples are obtained if accessions are further apart and smaller samples if accessions are closer together. Genetic distance sampling can be used in conjunction with predefined stratifications of the accessions. Sample sizes are determined automatically; they depend on the distances between accessions within strata. The method is applied to the collection of cultivated lettuce of the Centre for Genetic Resources, the Netherlands. In this paper, genetic distances between accessions are obtained using AFLP marker data. However, genetic distance sampling can be applied using any measure of genetic distance between accessions. Some properties of genetic distance sampling are discussed.

Communicated by A. Charcosset.

J. Jansen (✉)
Biometris, Wageningen University and Research Centre,
P.O. Box 100, 6700 AC Wageningen, The Netherlands
e-mail: johannes.jansen@wur.nl

Th.van Hintum
Centre for Genetic Resources, the Netherlands (CGN),
Wageningen University and Research Centre, P.O. Box 16,
6700 AA Wageningen, The Netherlands

Introduction

Gene banks have been founded with the aim to conserve the genetic diversity of crop species. This diversity forms the raw material of plant breeding. If possible, genetic diversity, also referred to as germplasm, is conserved in the form of accessions: batches of seed sampled from wild populations, traditional landraces, modern cultivars, genetic stock or other research material.

Many gene banks currently face problems caused by the large sizes of collections, and the resulting costs of maintaining these collections. This may endanger the long-term conservation of the collections. In addition, excessive collection sizes may hinder the accessibility by the users of genetic diversity, such as plant breeders (van Hintum et al. 2000). The concept of core collections was introduced by Frankel (1984). A core collection is a collection of limited size with the aim to represent the genetic diversity (or spectrum) of the whole collection (Brown 1995). From this definition of a core collection it follows that it should be avoided that not only identical accessions but also similar (or near-identical) accessions become part of a core collection.

Several methods have been introduced for sampling accessions from a gene bank collection to form a core collection. These methods include simple random sampling and stratified random sampling, but also more sophisticated methods. Schoen and Brown (1993, 1995) describe a method (referred to as M strategy) by which entries of the core collection are selected by minimizing the overall probability that an allele present in the gene bank collection is not retained in the core collection. The computer program MSTRAT

(Gouesnard et al. 2001) performs a generalized form of the M strategy. In the case a gene bank collection has been or can be divided in clearly distinct groups, stratified random sampling should be the method of choice (Brown 1989). With regard to stratified random sampling sample sizes may be obtained using the constant, the proportional or the logarithmic proportional method (Brown 1989). Brown (1989) provides a genetical justification for using the logarithmic proportional method. In the above, stratification does not involve information about the diversity between and within strata. Schoen and Brown (1993, 1995) describe a method (referred to as H strategy) by which sample sizes are obtained by maximizing the expected number of alleles retained in the core collection. Marita et al. (2000) describe an algorithm to identify accessions that are maximally diverse. A method for determining sample sizes based on genetic distances was introduced by Franco et al. (2005). Still, a major drawback of random sampling is that it cannot prevent that similar (or even identical) accessions are sampled from a gene bank collection.

Recently, molecular genetic markers have been used to characterize gene bank collections (Bretting and Widrlechner 1995; van Hintum and van Treuren 2002). Molecular genetic marker data can be used to calculate distances between accessions. These distances can be used to determine whether accessions are identical or similar. However, if accessions are identical or similar the problem remains which of the accessions should be chosen as entries of the core collection. If no additional information is available to support the choice of specific accessions, a form of random sampling should be retained in order to sample entries of the core collection. In this paper, a novel method called genetic distance sampling will be introduced.

Genetic distance sampling combines random sampling with information about genetic distances. The basic idea of the method is to start by sampling one accession at random and by discarding all accessions within a given distance to the sampled accession. This distance will be referred to as sampling radius. The process is continued by randomly sampling an accession from the remaining accessions and by discarding again all accessions within a given distance to the sampled accession. This process is continued until the set of remaining accessions is empty. If the sampling radius is set equal to 0, only duplicates with regard to the marker information will be removed. A major effect of genetic distance sampling is that clusters of accessions are automatically represented by one accession, or perhaps a few depending on the distances within the clusters. The method automatically adapts sample sizes depending on the genetic distances within clusters.

The AFLP marker data used in this study were generated in a much larger project aimed at characterizing the entire lettuce collection of the Centre for Genetic Resources, the Netherlands (CGN) with molecular markers (van Hintum 2003).

Materials and methods

Data

A detailed description of the plant material, the DNA extraction techniques and the AFLP analysis can be found in Jansen et al. (2006). In this study, 1,287 accessions of *Lactuca sativa* from the CGN collection are used. The CGN collection of *L. sativa* has been divided into seven lettuce types: butterhead (represented by 668 accessions), cos (187), crisp (203), cutting (138), latin (54), stalk (27) and oilseed (6). Each accession is represented by a single series of binary observations on 149 polymorphic AFLP markers. For all accessions included in this study at least 90% of the marker observations are present. The AFLP observations on any pair of accessions can be represented by means of a 3×3 contingency table (Table 1).

Genetic distance sampling

Genetic distance

The genetic distance between two accessions used in this paper, can be written as

$$\Delta = \frac{n_{01} + n_{10}}{n_{00} + n_{01} + n_{10} + n_{11}}.$$

The quantity $1 - \Delta$ is known as the simple matching coefficient. An alternative, simple measure of genetic distance is obtained by replacing the simple matching coefficient with Jaccard's similarity coefficient. A general treatment of distance measures is given by Gower (1971). A comparison of distance measures for

Table 1 Summary of observations on two accessions

Number of individuals		Accession 2		
		No band	Band	Missing
Accession 1	No band	n_{00}	n_{01}	n_{0u}
	Band	n_{10}	n_{11}	n_{1u}
	Missing	n_{u0}	n_{u1}	n_{uu}

genetic distance sampling goes beyond the scope of this paper.

Genetic distance sampling

The starting point is a list of all accessions $\mathbf{A}_1$. Distances between accessions are assumed to be known. Accessions that will be assigned to the core are sampled one at a time.

The first entry of the core is sampled at random from $\mathbf{A}_1$, and is called E_1. Accessions with a distance to E_1 smaller than the selection radius r are discarded from the list of accessions $\mathbf{A}_1$; the new list of accessions is called $\mathbf{A}_2$. The accessions that have been discarded are put in a list called $\mathbf{D}_1$. The second entry of the core is sampled at random from $\mathbf{A}_2$, and is called E_2. Accessions with a distance to E_2 smaller than the selection radius r are discarded from the list of accessions $\mathbf{A}_2$; the new list of accessions is called $\mathbf{A}_3$. The accessions that have been discarded are put in a list called $\mathbf{D}_2$. This process is repeated until the list of accessions becomes empty.

The above-described procedure leads to a list of entries of the core, $E_1, E_2,...,E_S$, in which S is the size of the core. The size of the core S depends mainly on the sampling radius r. Due to the nature of the sampling procedure the value of S and the composition of the final core collections will usually vary if the sampling procedure is repeated. The procedure also leads to a list of lists of accessions $\mathbf{D}_1$, $\mathbf{D}_2,...,\mathbf{D}_S$, which contain accessions within a distance r to E_1, $E_2,...,E_S$, respectively.

Stratified genetic distance sampling

Accessions may have been grouped into accessions of a different nature, e.g. accessions from different geographical regions may have been put in different groups. A stratification of accessions can be taken care of by adding an extra restriction to the distance restriction. The elements of $\mathbf{D}_s$ should not only be within a distance r from E_s, they should also be of the same stratum as $E_s (s = 1, 2,...,S)$. In this way, more than one stratification can be applied simultaneously, e.g. a two-way classification of accessions. Application of genetic distance sampling ensures that in the core each stratum is represented by at least one accession. The actual numbers of individuals that are sampled for different groups depend on the genetic distances within those groups. Genetic distance sampling will automatically sample more individuals from groups with larger within-group distances than from groups with smaller within-group distances.

Computations

Computations have been carried out using the statistical package Genstat (Genstat Committee 2005) and special purpose programmes written in C (Kernighan and Ritchie 1988).

Results

Relationship between sample size and selection radius

Initially, no distinction was made between the seven lettuce types. Figure 1 shows the relationship between the size of the core obtained using genetic distance sampling, and the selection radius r. Figure 1 shows that only small differences in sample size occur between runs with the same value of the selection radius. By setting r to 0, it was found that only 1,117 distinct 'AFLP profiles' (87%) are present in the collection of 1,283 accessions of cultivated lettuce, using 149 polymorphic markers.

Comparison of genetic distance sampling with stratified random sampling

As an example, Table 2 shows core sizes obtained with genetic distance sampling and with stratified random

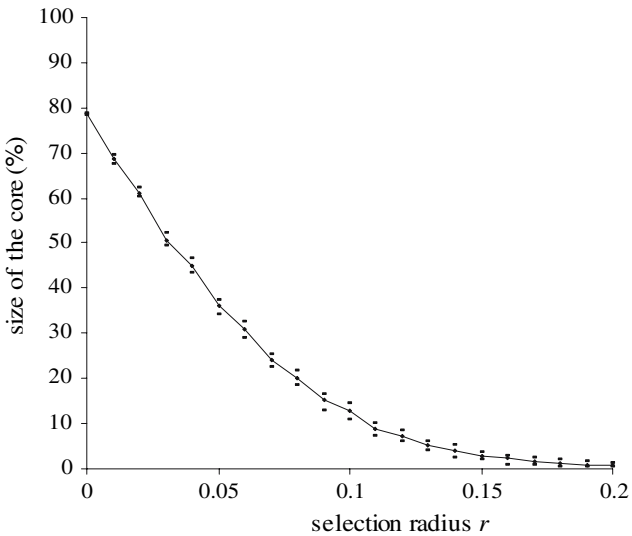

Fig. 1 Relationship between the sample size (as percentage of the entire collection) and the selection radius r for stratified genetic distance sampling as obtained for the lettuce data. The *solid curve* shows the average core size of 1,000 runs of stratified genetic distance sampling. *Dashes* indicate minimum and maximum core sizes

Table 2 An example of core sizes obtained using genetic distance sampling with selection radius 0.108, using stratified random sampling (proportional and logarithmic strategy) and using stratified genetic distance sampling with selection radius 0.108

Type	Collection	Number of accessions in core			
	Number of accessions	Genetic distance sampling, r = 0.108 (example)	Stratified random sampling (P)	Stratified random sampling (L)	Stratified genetic distance sampling, r = 0.108 (example)
Butterhead	668	30	66	27	32
Cos	187	36	19	22	43
Crisp	203	15	20	22	21
Cutting	138	31	14	20	41
Latin	54	8	5	17	17
Stalk	27	6	3	14	9
Oilseed	6	2	1	6	3
Total	1,283	128	128	128	166

For the stratified random sampling approaches, the total number of accessions in the core was set to 10% of the total number of accessions in the entire collection

sampling with a total sample size of 128 (10% of the entire collection). For stratified random sampling the proportionality rule and logarithmic proportionality rule are used. Major differences occur between the methods. First, genetic distance sampling and stratified random sampling according to the proportionality rule are considered. For the large groups, butterhead and crisp, genetic distance sampling produces core sizes which are smaller than those obtained with stratified random sampling. For the other lettuce groups, genetic distance sampling produces cores with are larger than those obtained with stratified random sampling using the proportionality rule.

Compared to genetic distance sampling and stratified random sampling according to the proportionality rule, stratified random sampling using the logarithmic proportionality rule assigns comparatively large core sizes to the lettuce types with a small number of accessions in the collection (latin, stalk and oilseed).

Sampling within groups: stratified genetic distance sampling

It is also possible to carry out sampling individuals within lettuce types. Table 2 shows example results of stratified genetic distance sampling. In this case, stratified genetic distance sampling produces a larger core than simple genetic distance sampling. This is due to the overlap between lettuce types and the additional restriction that sampled accessions and associated discarded accessions should be of the same lettuce type. In Fig. 2, the sampled proportions of accessions for the seven lettuce types have been plotted against the corresponding average distances between accessions within lettuce types. In Fig. 2, the value of r was set to 0.125, which provides total sample sizes that are on average slightly smaller than 10% of the entire collection. Figure 2 shows that within the range of values the sampling proportions are approximately linear with the average distances between accessions within lettuce types.

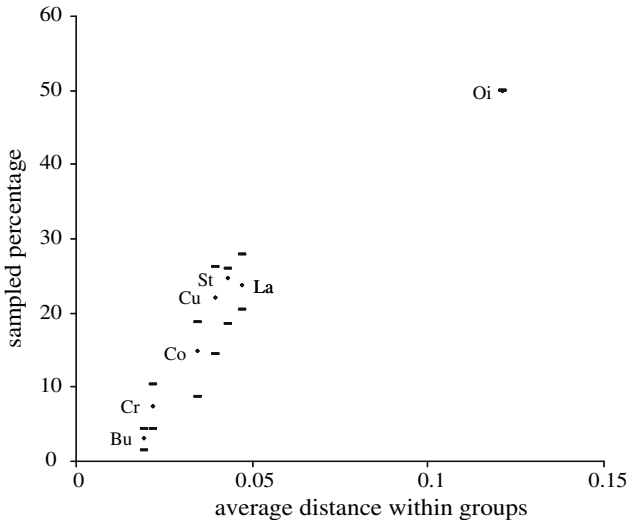

Fig. 2 Average percentages of sampled accessions for the seven lettuce types obtained using stratified genetic distance sampling (r set to 0.125) plotted against average distances between accessions within lettuce types in the entire collection. Average percentages of sampled accessions are based on 1,000 samples. *Dashes* indicate minimum and maximum percentages of sampled accessions. *Bu* butterhead, *Co* cos, *Cr* crisp, *Cu* cutting, *La* latin, *St* stalk

Band frequencies

For (stratified) random sampling procedures, the band frequencies of AFLP markers in the sample are unbiased estimates of the band frequencies of AFLP

markers in the entire collection. For random sampling procedures, all accessions that have not been sampled in the current round of sampling have a probability equal to 1 divided by the remaining number of accessions of being sampled in the next round. However, for genetic distance sampling all accessions that have not been sampled until the current round of sampling have either probability 0 of being sampled in the next round of sampling (i.e. if they are within a distance r of an individual sampled already) or a probability equal to 1 divided by the remaining number of accessions of being sampled in the next round of sampling. As a consequence, genetic distance sampling may provide biased estimates of the band frequencies of the AFLP markers.

For six lettuce types, Fig. 3 shows the averages of the band frequencies of the AFLP markers in the sample obtained using 1,000 runs of stratified genetic distance sampling versus the band frequencies of the AFLP markers in the entire collection. For random sampling procedures, the averages based on 1,000 runs are very close to the line of equality (results not shown). For stratified genetic distance sampling, the averages deviate considerably from the line of equality, in a systematic as well as in a random manner. The dashed curve represents a third-degree polynomial, which is forced through the points (0,0), (0.5,0.5) and (1,1). The formula for this polynomial may be written as $y = x + \alpha(1/2x - 3/2x^2 + x^3)$, in which y represents the average band frequency in the sample and x rep-

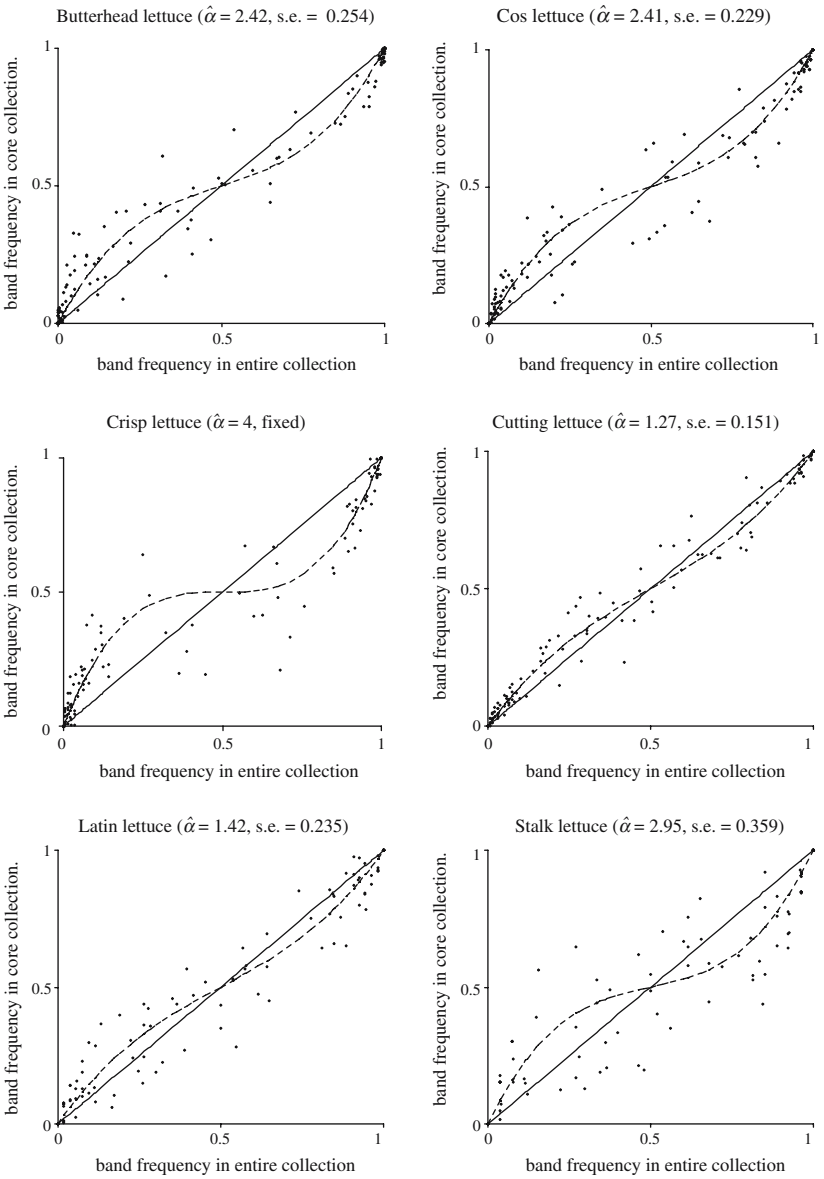

Fig. 3 Average band frequencies of AFLP markers in samples (y) obtained using stratified genetic distance sampling (r = 0.125) plotted against the corresponding band frequencies in the entire collection (x). The averages are based on 1,000 samples. The *solid lines* represent the line of equality. The *dashed curves* represent the line $y = x + \alpha(1/2x - 3/2x^2 + x^3)$. Estimates of â for the different lettuce types and the corresponding standard errors (s.e.) have been obtained using linear regression (using x as an offset variable)

resents the band frequency in the entire collection. For $\alpha = 0$, the line of equality is obtained. For all lettuce types, the value of α (values shown in Fig. 3) has been found to be significantly greater than 0. As a consequence, for AFLP markers with a band frequency close to 0 (1) in the entire collection, the band frequency is on average increased (decreased) in the sample.

The above results may also have major consequences for the probability that an AFLP marker which is polymorphic in the entire collection becomes non-polymorphic in the sample. This probability will be denoted by the acronym PLP (probability of loss of polymorphism). Figure 4 shows the results of a comparison of stratified random sampling (L strategy) and genetic distance sampling. It may be concluded from Fig. 4 that for the crop types butterhead, cos, crisp and cutting the values of PLP obtained for stratified genetic distance sampling are in general much smaller than corresponding values of PLP obtained for stratified random sampling (L strategy). For lettuce types, latin and stalk the advantage of stratified genetic distance sampling over stratified random sampling (L strategy) is much smaller.

Discussion

(Stratified) genetic distance sampling provides an efficient procedure for incorporating distances between accessions into a random sampling framework. It avoids the tedious computations that are required by optimization procedures and is intuitively clear. Ge-

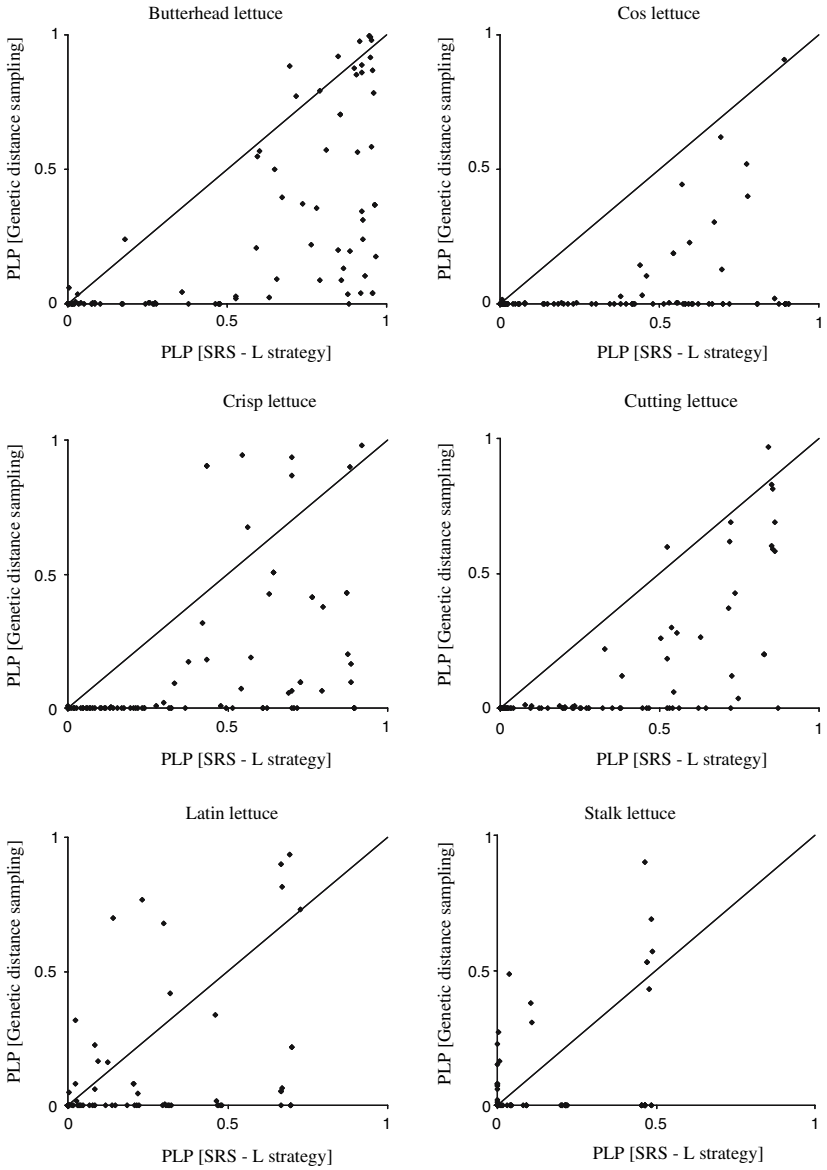

Fig. 4 Probability of loss of polymorphism for stratified genetic distance sampling with r set to 0.125 plotted against PLP for stratified random sampling (L strategy; sample sizes shown in Table 2). Values have been obtained using 1,000 samples for each of the sampling methods

netic distance sampling still requires specially written software. Distance information can be used in conjunction with simple, random sampling. It can also be incorporated into stratified, random sampling by putting a further restriction on accessions to be discarded from sampling, namely that they should not only be within the sampling radius of a sampled accession but also of the same type (stratum) as the sampled accession. Other restrictions on the sampling would also be possible, such as geographical restrictions. For example, accessions can only be discarded if they have been collected within a certain geographical distance from the sampled accession they are genetically associated with (Charmet and Balfourier 1995).

A major advantage of genetic distance sampling is that a relatively small number of accessions are sampled from groups that are homogeneous, and that a relatively large number of accessions are sampled from groups that are heterogeneous. The same idea is used by Franco et al. (2005). Their D allocation determines the size of the sample drawn from a cluster to be proportional to the mean distance between individuals within that cluster. Genetic distance sampling does not require prior determination of clusters. This avoids arbitrary decisions involved in clustering methods; for example, decisions about how to define distances between groups.

The core size obtained by genetic distance sampling depends on the selection radius r. Figure 1 shows that very little variation about the relationship between the core size S and the sampling radius r is present. As a consequence, a value of r, for which $S(r) \approx S_0$, in which S_0 is the required size of the core, can be obtained by simple optimization techniques, or simply by increasing (decreasing) the value of r if the core size becomes too large (small).

Genetic distance sampling does not only determine a core collection. Each accession not included in the core is associated with an accession in the core, the distance between the two accessions being smaller than the sampling radius r. Due to the random nature of the sampling procedure accessions not included in the core are not necessarily associated with the nearest accession in the core. The outcome of the sampling procedure can be further improved by determining for each accession not included in the core the nearest accession in the core (taking the stratification of the accessions into account). The association of accessions in the collection to an accession in the core may assist the user of gene banks in finding alternatives for accessions in the core collection, or in extrapolating knowledge about accessions in the core to accessions in the collection.

Instead of, or in addition to, the AFLP fingerprints used in this paper, other of variables of a different nature (qualitative and quantitative measurements) can be integrated in a distance measure (Gower 1971). This allows a more general applicability of genetic distance sampling than to molecular marker data, enabling the creation of core collections based on very basic characterization information in conjunction with a stratification based on passport data.

Schoen and Brown (1993, 1995) proposed methods (H strategy, M strategy) for obtaining core collections of fixed size using optimization. In this case, these approaches tend to obtain samples of accessions for which the band frequencies of AFLP markers are pushed away from the boundaries 0 and 1. This appears also the case with (stratified) genetic distance sampling, but not with random sampling strategies.

So far, genetic distance sampling has only been applied using AFLP marker data obtained in a gene bank collection of cultivated lettuce. Full-scale application of genetic distance sampling would require successful applications on data from several plant collections using data of various types (marker data, phenotypic data). A further study will include a comparison of genetic distance sampling and (stratified) random sampling and also a comparison various distance measures. Cross-validation will be a useful tool for investigating the efficiency of genetic distance sampling with regard to capturing genetic variation. Successful practical application of genetic distance sampling also requires easily accessible computer software.

Acknowledgments The authors are very grateful to the editor and two referees, whose comments led to various improvements of the paper.

Appendix

In order to provide a mathematical description of the relationship between the average band frequencies of AFLP markers in samples obtained using genetic distance sampling (y) and the corresponding band frequencies in the entire collection (x) a third-order polynomial was used:

$$y = f(x) = \alpha x^3 + \beta x^2 + \gamma x + \delta.$$

This function provides enough flexibility to describe the relationship between y and x. Since the band frequency of non-polymorphic AFLP markers remain unchanged under any sampling procedure, $f(0) = 0$, leading to $\delta = 0$, and $f(1) = 1$, leading to $\gamma = 1 - \alpha - \beta$. As a consequence,

$$f(x) = a(x^3 - x) + \beta(x^2 - x) + x. \quad (1)$$

Since the simple matching coefficient treats the presence of bands in the same way as the absence of bands, it follows that $f(1/2) = 1/2$, leading to $\beta = -3/2\ \alpha$. As a consequence,

$$f(x) = \alpha\left(x^3 - \frac{3}{2}x^2 + \frac{1}{2}x\right) + x. \quad (2)$$

In order to achieve that the function $f(x)$ is non-decreasing, the value of α should be smaller or equal to 4. If α is positive (negative), the slope of $f(x)$ is greater (smaller) than unity if x is either close to 0 or 1.

Using the least-squares criterion, the function $g(x) = f(x) - x$ can be fitted to the data by simple linear regression with zero intercept. For the data used in this study expression (1) did not provide a significantly better fit to the data compared to expression (2). Therefore, only results using expression (2) have been presented. It would also be possible to use weighted linear regression with weights proportional to $1/x(1 - x)$. This would give more weight to points with x close to 0 or 1 in comparison to points with x close to 1/2. For the current data this leads to even larger estimates of α.

References

Bretting PK, Widrlechner MP (1995) Genetic markers and plant genetic resource management. Plant Breed Rev 31:11–86

Brown AHD (1989) Core collections: a practical approach to genetic resources management. Genome 31:818–823

Brown AHD (1995) The core collection at the crossroads. In: Hodgkin T, Brown AHD, van Hintum ThJL, Morales EAV (eds) Core collections of plant genetic resources. Wiley, Chichester, pp 3–19

Brown AHD, Schoen DJ (1994) Optimal strategies for core collections of plant genetic resources. In: Loeschcke V, Tomiuk J, Jain SK (eds) Conservation genetics. Birkhäuser, Basel, pp 357–370

Charmet G, Balfourier F (1995) The use of geostatistics for sampling a core collection of perennial ryegrass populations. Genet Resour Crop Evol 42:303–309

Franco J, Crossa J, Taba S, Shands H (2005) A sampling strategy for conserving genetic diversity when forming core subsets. Crop Sci 45:1035–1044

Frankel OH (1984) Genetic perspectives of germplasm conservation. In: Arber W, Llimensee K, Peacock WJ, Starlinger P (eds) Genetic manipulation: impact on man and society. Cambridge University Press, Cambridge, pp 161–170

Genstat Committee (2005) Genstat® Release 8. Reference manual. VSN International, Hemel Hempstead

Gouesnard B, Bataillon TM, Decoux G, Rozale C, Schoen DJ, David JL (2001) MSTRAT: an algorithm for building germplasm core collections by maximizing allelic or phenotypic richness. J Hered 92:93–94

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857–871

van Hintum ThJL (2003) Molecular characterisation of a lettuce germplasm collection. In: Eucarpia leafy vegetables 2003, Proceedings of the Eucarpia meeting on leafy vegetables genetics and breeding, Noordwijkerhout, The Netherlands, 19–21 March 2003. Centre for Genetic Resources, Wageningen, pp 99–104

van Hintum ThJL van Treuren R (2002) Molecular markers: tools to improve genebank efficiency. Cell Mol Biol Lett 7:737–744

van Hintum ThJL, Brown AHD, Spillane C, Hodgkin T (2000) Core collections of plant genetic resources. International Plant Genetic Resources Institute, Rome

Jansen J, Verbakel H, Peleman J, van Hintum ThJL (2006) A note on the measurement of genetic diversity within genebank accessions of lettuce (*Lactuca sativa* L.) using AFLP markers. Theor Appl Genet (in press)

Marita JM, Rodriguez JM, Nienhuis J (2000) Development of an algorithm identifying maximally diverse core collections. Genet Resour Crop Evol 47:515–526

Schoen DJ, Brown AHD (1993) Conservation of allelic richness in wild crop relatives is aided by assessment of genetic markers. Proc Natl Acad Sci USA 90:10623–10627

Schoen DJ, Brown AHD (1995) Maximising genetic diversity in core collections of wild relatives of crop species. In: Hodgkin T, Brown AHD, van Hintum ThJL, Morales EAV (eds) Core collections of plant genetic resources. Wiley, Chichester, pp 55–76

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414