Second RUFORUM Biennial Meeting 20 - 24 September 2010, Entebbe, Uganda

Research Application Summary

Determination of genetic structure of germplasm collections: Are traditional hierarchical clustering methods appropriate for genetic marker data?

Odong, T.L.[1,2],van Heerwaarden, J.[1], Jansen, J.[1], van Hintum, T.J.L.[3] & van Eeuwijk, F.A.[1]
[1]Biometrics, Wageningen University UR, Wageningen, The Netherlands
[2]Department of Crop Science, Faculty of Agriculture, Makerere University,
P. O. Box 7062, Kampala, Uganda
[3]Centres for Genetic Resources, The Netherland, Wageningen, The Netherlands
Corresponding author: thomas.odong@wur.nl

Abstract

We studied the performance of traditional hierarchical clustering techniques using molecular marker data. In this study, we showed that the cophenetic correlation coefficient is directly related to subgroup differentiation and can thus be used as an indicator of the presence of genetically distinct subgroups in germplasm collections. Whereas UPGMA performed well in preserving distances between accessions, Ward excelled in recovering groups. Our results also showed a close similarity between clusters obtained by Ward and by model-based cluster method (STRUCTURE). Traditional cluster analysis can provide an easy and effective way of determining structure in germplasm collections using molecular marker data.

Key words: Cophenetic correlation coefficient, Subgroup differentiation, STRUCTURE, UPGMA, Ward

Résumé

Nous avons étudié la performance des techniques traditionnelles de classification hiérarchique en utilisant des données de marqueurs moléculaires. Dans cette étude, nous avons montré que le coefficient de corrélation cophénétique est directement lié à la différenciation des sous-groupes et peut donc être utilisé comme un indicateur de la présence de sous-groupes génétiquement distincts dans les collections du matériel génétique. Tandis qu'UPGMA a donné de bons résultats dans la préservation de distances entre les accessions, Ward a excellé dans la récupération des groupes. Nos résultats ont également montré une similitude étroite entre les groupes obtenue par Ward et par la méthode de groupe basée sur un modèle (STRUCTURE). L'analyse traditionnelle de groupes peut constituer un moyen facile et efficace de détermination de la structure dans les collections du matériel génétique à partir des données de marqueurs moléculaires.

Mots clés: Coefficient de corrélation cophénétique, la différenciation des sous-groupes, STRUCTURE, UPGMA, Ward

Background

Determination of the genetic structure of germplasm collections is of fundamental importance for both conservation and utilization of genetic resources assembled in genebanks and other research facilities around the world. With large numbers of accessions being accumulated in genebanks and other research facilities, curators are faced with numerous choices on how best to conserve these resources and at the same time make them available to those who would like to utilize them in breeding programs. The approach of forming core collections was introduced to increase the efficiency of characterization and utilization of collections stored in genebanks, while preserving as much as possible the genetic diversity of the entire collection (Brown, 1989). A core collection provides a structured sample from the entire collection, one that is more manageable in size than the whole collection. One of the key problems in forming core collections is how to partition the heterogeneous collections into more homogeneous sub-groups within which sampling can be performed. The partitioning of heterogeneous germplasm collections into homogeneous sub-groups is very closely related to the identification of population substructure in association panels and forms an essential component of association mapping studies.

Although many new approaches for determining population genetic structure have been introduced, traditional hierarchical clustering is still very popular. However, little is known about the appropriateness of traditional hierarchical clustering methods for recovering population genetic structure contained in molecular marker data.

Literature Summary

In genebanks, collections are organised in different ways, but they all consist of some or all of the following: landraces and selected lines from landraces, elite breeding lines, released varieties, wild and weedy relatives of the cultigen, and genetic stocks most often from different areas of origin (Vaughan and Jackson, 1995). It is clear that the genetic diversity of crop species in genebanks and other research facilities has originated in a way totally different from that of natural populations. Therefore the appropriateness of new approaches for determination of population genetic structure such as STRUCTURE (Pritchard *et al*., 2000) which are based on the assumptions of natural populations may be questionable. The traditional hierarchical clustering methods on the other hand do not require population genetic assumptions such as Hardy-Weinberg or linkage disequilibrium. However, very few

evaluations of hierarchical clustering have been done using molecular marker data. Most studies have been performed using quantitative variables such as plant height, yield etc. In addition, most of the datasets considered were of limited sizes both in terms of number of objects and variables (Milligan and Cooper, 1985).

Study Description

We studied the relationship between population genetic structures as measured by subgroup differentiation (F_{ST}) and dendrogram evaluation criteria, and the ability of the clustering techniques to identify and recover groups in the data using real and simulated data sets.

Three real data sets (coconut, potato and common beans) used in this study were generated under the auspices of the Generation Challenged Programme-GCP (www.generationcp.org). The coconut data set consisted of 1014 accessions genotyped with 30 Simple Sequence Repeat (SSR) markers, the potato data set consisted of 230 diploid accessions genotyped with 50 SSR markers and the common beans data set consisted of 636 accessions genotyped with 33 SSR markers. The accessions studied originated from different regions of the world. For simulated data, genotypic data were generated by SimuPOP (Peng and Kimmel, 2005), a forward-time population genetic simulation environment. We used a "Finite Island" as well as "Stepping Stone Migration" (Kimura, 1953) models.

For the purpose of this presentation, we have categorized the evaluations of hierarchical cluster analysis results into: 1) measurement of the agreement between input distance matrices and distances as fitted in dendrograms, and 2) quantification of the amount of structure in the data set. The cophenetic correlation coefficient (CPCC) (Sokal and Rohlf, 1962) was used to measure agreement between input distance matrix and the results of a hierarchical cluster analysis while agglomerative coefficient (AC) (Kaufman and Rousseeuw, 1990) was used to quantify the amount of structure present in the data.

For each of the three real data sets, clustering was performed using UPGMAand Ward's method. CPCC and AC were calculated for each dendrogram. For real data sets (out crossing species: coconut and potato) the results of cluster analysis using UPGMA and Ward's method were also compared with the output from model-based clustering method, STRUCTURE.

To explore the relationships between F_{ST} and evaluation criteria, datasets from different simulations were pooled together and then grouped based on the strength of subgroup differentiation into groups (each containing 100 datasets) with similar realized values of F_{ST}. For each dataset, cluster analysis was performed using UPGMA and Ward methods. For each dendrogram constructed, both CPCC and AC were calculated. Hierarchical cluster analysis was performed using Agglomerative Nesting (Agnes) procedure (Kaufman and Rousseeuw, 1990) in the package Cluster of R. The ability of UPGMA and Ward's method to recover the subpopulations in the data was evaluated using overall cluster purity (maximum proportion of cluster members coming from a single subpopulation).

Research Application

We showed that the cophenetic correlation coefficient (CPCC)) is directly related to F_{ST} and can therefore be used as an indicator of the presence of genetically distinct subgroups in germplasm collections. Larger CPCC ($- \geq 0.8$) with both UPGMA and Ward's method is an indication of a good subgroup differentiation in the data. The use of AC for determination of genetic structure in a data set does not give reliable results. While UPGMA performed much better in preserving the original pair wise relationships between accessions, Ward's method excelled in recovering the original groups in the data. Our results also showed a good similarity between clusters formed by Ward's method, model-based clustering method (STRUCTURE) and passport data. It was clear from our studies that because of the unique characteristics of accessions in genebanks, very often application of standard criteria for evaluating the quality of dendrograms may not work. In conclusion, traditional cluster analysis can provide an easy and in some cases an effective way of determining structure in germplasm collections.

≥ 0.8

Recommendation

In a situation in which there is a clear indication of the presence of genetically distinct subgroups (i.e., when CPCC- with both UPGMA and Ward's dendrograms), groups formed by Ward's method should be used. Groups formed by UPGMA should be used in a situation in which only UPGMA produces dendrogram with CPCC. Results obtained from the traditional clustering techniques should be compared with results obtained from other methods such as passport data.

References

Brown, A.H.D. 1989. Core collections - A practical approach to genetic-resources management. *Genome* 31(2):818-824.

Kaufman, L. and Rousseeuw, P.J. 1990. Finding groups in data. An introduction to cluster analysis. Wiley-Interscience, New York.

Kimura, M. 1953. Stepping stone model of population. *Ann. Rept. Nat. Inst. Genetics* 3:62-63

Milligan, G.W. and Cooper, M.C. 1985. An examination of procedures for determining the number of clusters. In: A data set. *Psychometrika* 50(2):159-179.

Peng, B. and Kimmel, M. 2005. SimuPOP: A forward-time population genetics simulation environment. *Bioinformatics* 21(18):3686-3687.

Pritchard, J.K. and Stephens, M. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155(2): 945-959.

Sokal, R.R. and Rohlf, F.J. 1962. The comparison of dendrograms by objective methods. *Taxon* 11:33 - 40.

Vaughan, D.A. and Jackson, M.T.1995. The core as a guide to the whole collection. In: Hodgkin, T., Brown, A.H.D., Th. J.L. van Hintum and Morales, E.A.V. (Eds.). Core collections of plant genetic resources. John Wiley & Sons, Chichester, pp. 229-239.

Wang, W.Y.S., Barrat, B.J., Clayton, G.G. and Todd, J.A. 2005. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* 6:109 - 118.

Ward, J.H. 1963. Hierarchical groupings to optimize an objective function. *Journal of the American Statistical Association* 58:236 - 244.