Unbiased estimation of pairwise genetic similarity from binary AFLP data

Gerrit Gort and Fred van Eeuwijk[1]

[1] Biometris, Wageningen University, The Netherlands

AFLP is a DNA fingerprinting technique frequently used in the plant sciences. The resulting profile for an individual plant is a sort of barcode of bands, resulting from the separation of DNA fragments by length on an electrophoretic gel. This barcode is often represented as a binary vector, indicating absence or presence of bands. AFLP is used e.g. in phylogenetic studies to estimate the pairwise genetic similarity of individuals (from different species or accessions). We interpret pairwise genetic similarity as the fraction of DNA shared by two individuals, amounting to the fraction of shared DNA fragments in the case of the AFLP procedure. This genetic similarity is usually estimated by the Dice (or Jaccard) coefficient, calculated from the pair of binary band vectors of the two individuals.
In this study we focus on two problems in the interpretation of AFLP profiles. 1) Within a profile two or more fragments of the same length but of different genomic origin may have been formed, which collide into a single band. We call this collision, see Gort (2006) and Gort (2008). 2) In a pair of profiles two fragments of the same length but, again, of different genomic origin may have been formed, appearing as identical bands in the two profiles. This problem is called homoplasy. As a consequence similarity coefficients calculated from the binary band information overestimate the true genetic similarity, with increasing bias for higher numbers of bands per profile. Koopman and Gort (2004) studied the problem, calculating similarities in case of genetic similarity equal to 0.
We now propose two different estimators of pairwise genetic similarity. In the first type the band counts in the Dice coefficient are simply replaced by estimated fragment counts. This works for cases where band lengths are known or unknown. Precisions are estimated using bootstrapping. In the second type the genetic similarity is estimated by maximum likelihood, and precision follows from standard likelihood theory. For this estimator known band lengths are needed. Characteristics of the estimators are studied by simulation.
We use data from a study on lettuce as an example.

References

[1] Gort G, Koopman WJM, Stein A, vanEeuwijk FA (2008) Collision probabilities for AFLP bands, with an application to simple measures of genetic similarity. JABES to appear

[2] Gort G, Koopman WJM, Stein A (2006) Fragment length distributions and collisions probabilities for AFLP markers. Biometrics 62: 1107-1115

[3] Koopman WJM, Gort G (2004) Significance tests and weighted values for AFLP similarities based on *Arabidopsis* in silico AFLP fragment length distributions. Genetics 167: 1915-1928