ACCOUNTING FOR DEPENDENCE IN SIMILARITY DATA FROM DNA FINGERPRINTING

G. Hepworth[†], I.R. Gordon, M.J. McCullough

The University of Melbourne, Melbourne, Australia

[†] E-mail: *hepworth@unimelb.edu.au*

It is of interest to molecular biologists to assess the genetic similarity of strains of a pathogen, such as the yeast *Candida*, which causes thrush in humans. If a group of strains with a particular attribute (such as country of origin or the presence of HIV) is more similar than expected, this gives insight into the characteristics of the strains.

Molecular techniques such as PCR and RAPD can produce a DNA fingerprinting profile for each strain, consisting of bands scored as present or absent. The similarity of a group of strains may then be assessed by calculating a matrix of dissimilarities from these profiles. It is then possible to use the mean dissimilarity for each strain across other strains within the group as an observation in a statistical analysis. However, these observations are not independent of each other, so standard analysis techniques which ignore this dependence are inappropriate, because they underestimate the variance of the group means.

By examining the correlation between elements of the dissimilarity matrix, we show that the variance is underestimated by a factor of between about 2 and 4, which seriously affects inference and subsequent conclusions. Permutation tests provide an appropriate way of accounting for the dependence, and this is illustrated using a study involving strains of *Candida*.