# 2005 GBIF Ebbe Nielsen Prize:

## *NDM / VNDM: Computer Programs to Identify Areas of Endemism*

*Pablo A. Goloboff*

Instituto Superior de Entomología (Facultad de Ciencias Naturales, Universidad de Tucumán), and Consejo Nacional de Investigaciones Científicas y Técnicas, Miguel Lillo 205, 4000 S.M. de Tucumán, Argentina
pablogolo@csnat.unt.edu.ar

A taxon is said to be endemic to an area when it is found in that area but nowhere else. If limits in distribution of multiple, different groups of plants and animals are determined by the same factors, it is expected that the distributions of those groups will show similar patterns. Therefore, it is also expected that some areas or regions – so-called areas of endemism – will consistently house taxa that represent multiple different groups that are found only there and nowhere else. Although the notion of endemism is an old one that dates back to the early 1800s, attempts to identify areas of endemism by formal, quantitative means have begun only recently. The most widely used method records presence/absence data in a grid, and subjects the resulting data matrix to analysis by algorithms that implement criteria established for reconstructing phylogeny. However, these criteria are problematic when used to try to determine areas of endemism. Recent publications by this author and his collaborator Claudia Szumik have proposed alternative criteria that are designed specifically for recognition of areas of endemism. These criteria are implemented in two interacting computer programs, NDM and VNDM. These programs take as input vouchered locality data taken from natural history specimens (such as those made available by GBIF), and convert these into a grid of presence/absence data (optionally, probable presence can be used as a third category). The criterion used is very simple: for a given set of cells (or "area"), a given taxon can be considered as "endemic" to that area if it is more or less evenly distributed across the area, but is not found in distant cells. Thus, choosing sets of cells for which many taxa are endemic provides a natural way to identify areas of endemism. However, because the program has to examine, in principle, all possible combinations of cells, a computational problem arises. For a grid of *c* columns and *r* rows, there are

$$\sum_{i=2}^{c.r-1} \frac{(c.r)!}{i!\,(c.r-i)!}$$

possible combinations of cells. Even for a modest $c = 10$, $r = 20$, there are more than $22 \times 10^9$ possible sets of cells. Therefore, it is impossible to actually examine every possible cell combination when given input of the larger datasets that more realistically reflect biological reality. For this reason, a trial-and-error technique that has proven useful in other kinds of analyses is employed by the program NDM. The procedure uses promising sets of cells as starting points, modifies them by deleting or adding cells according to a specified set of rules, re-evaluates the degree of endemicity, and keeps the sets of cells with highest endemicity scores while discarding those sets with low scores. Also, endemicity analyses present additional problems when partially overlapping areas, or the combination of several areas into one, are attempted. An analysis of a real data set is presented to illustrate the approaches used by NDM and VNDM to dealing with these challenges.