WILEY-BLACKWELL

Cladistics (2015) 1–5

Cladistics

10.1111/cla.12112

Higher taxa and the identification of areas of endemism

Claudia A. Szumik and Pablo A. Goloboff*

Consejo Nacional de Investigaciones Científicas y Técnicas, Unidad Ejecutora Lillo, Miguel Lillo 251, S.M. de Tucumán 4000, Argentina

Accepted 6 January 2015

Abstract

Quantitative analyses of areas of endemism have rarely considered higher taxa. This paper discusses aspects related to the use of higher taxa in the analysis of areas of endemism, and computer implementations. An example of the application of the method is provided, with a data set for Nearctic mammals, showing that some of the areas recognized by species-level taxa also adjust well to the distribution of other taxa of higher level (genera, monophyletic groups).

One of the important goals in quantitative historical biogeography is identifying areas of endemism: areas determined by the congruent distribution of taxa. As in many other aspects of quantitative biogeography, some elementary aspects of the problem of identifying areas of endemism have long been neglected. An example is in how distributional data (typically incomplete and not necessarily conforming exactly to an area of endemism) may lead to ambiguous definitions of areas. This will be the case when minor differences in the delimitation of an area allow considering similar numbers of species as endemic, i.e. when the concordance of species distribution is high but not exact. Trivial as this seems, the possibility and the consequences of ambiguity had never been seriously considered until the problem of endemism was approached from the point of view of strict optimality criteria (Szumik et al., 2002); a proposed solution to cope with the ambiguity inherent to biogeographical distributions is the use of "consensus" areas (Aagesen et al., 2013).

This paper brings attention to another aspect of the problem which is no less trivial, but in practice equally ignored in quantitative analyses. Typically, analyses of endemism using some formal method (Morrone, 1994; Hausdorf and Hennig, 2003; Szumik and Goloboff, 2004) use as data the distribution of species (e.g. Carine et al., 2009; Szumik et al., 2012; Guedes et al., 2014), but many areas of endemism can be characterized by higher taxa. This had long been obvious to biogeographers since the 19th century (e.g. Wallace, 1876: 105 states that North America can be characterized by 13 families or subfamilies of Vertebrata, and an even more important number of genera). But in actual quantitative methods and applications, the possibility that higher taxa (instead of species) are the units that characterize some areas has so far been neglected.

Consider Fig. 1 as an example. It shows data for six species, in three genera (genera *Aus*, *Bus*, and *Cus*). The distributions of the individual species are not particularly congruent; no two species share a similar distribution. However, things change when the distribution of the genera is considered. This is shown in Fig. 2. The distribution of the genus *Aus* is the sum (union) of the distribution of its constituent species, which is identical to the distribution of the species *Bus cus*. Likewise for the genus *Cus*. Thus, the area of endemism is characterized by three taxa: two genera (*Aus* and *Cus*) and a species (*Bus bus*). The argument remains the same if (say) some of the higher taxa have a level above genus; every node in the tree corresponds to a taxon (whether named or not).

Therefore, it is clear that the information on higher taxa, i.e. which species are more closely related to each other, should be used in analyses of endemism.

*Corresponding author:
E-mail address: pablogolo@yahoo.com.ar

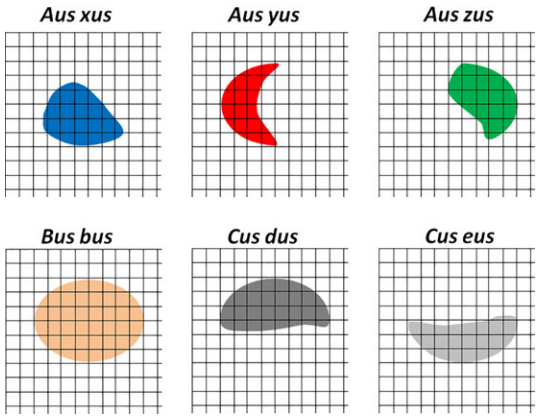

Fig. 1. Hypothetical example of species distributions, where no two species have congruent distributions.

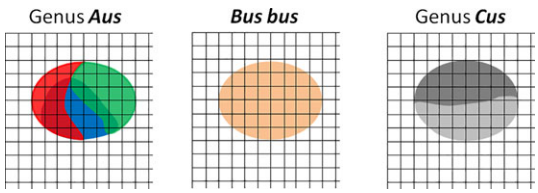

Fig. 2. When the species of Fig. 1 are grouped in genera, their distributions become congruent.

The problem is in how to incorporate that information into a particular quantitative analysis, where the areas that will result from the analysis are not known in advance. Simply replacing the distribution of the species in *Aus* and *Cus* by the distribution of their genera may lead to missing other smaller areas (characterized by some of those species). Adding the generic distributions to the species distributions may incorporate redundant information into the analysis—consider the case of a genus with two fully sympatric (co-distributed) species. The corresponding area cannot be considered as characterized at the same time by the two species and the genus, because that amounts to overcounting. As the two species in this hypothetical example have similar distributions, then they will have a similarly high degree of endemicity for the corresponding area (as measured, for example, by Szumik and Goloboff's, 2004 index). Thus, if we count the species instead of the genus, the score of endemicity (two taxa) for the area will be higher than if we count the genus (one taxon). Contrast this case with the one shown in Figs 1 and 2. For the genus *Aus* and the area shown in the figures, the score of endemicity will be higher if the genus (one taxon) is counted instead of the species (no endemic taxon).

This indicates the general rule for considering nested taxa in analyses of endemicity: if the endemicity score contributed by a higher taxon is more than the sum of the endemicity scores contributed by its subordinate taxa for the same area, then the higher taxon should be considered; otherwise, the subordinate taxa.

Simple as the rule is, applying it in real world analyses is difficult—there is no way to know in advance whether a higher taxon or a species must be counted. An analysis of endemicity should, ideally, consider all possible areas (cell combinations). Thus, the decision of considering the higher or the subordinate taxa will have to be made for each area scored during the analysis, on a per-area basis. Therefore, it is not possible for the user to create a data set that will automatically produce the desired effect; instead, the option to consider higher taxa must be incorporated into the program itself. Recent versions of VNDM, the program implementing Szumik and Goloboff's (2004) criterion (available at http://www.lillo.org.ar/phylogeny/endemism/VNDM-NDM_Nov_2014.zip), have incorporated this option. The program reads the distributions of the species and, optionally, a list of groups (higher taxa). Then, if groups have been defined, for each area to be scored for endemicity, the higher or the subordinate taxa are used, depending on which one produces the highest sum of scores.

Implementation

In VNDM, the definition of groups must be done in the same file containing the data. Groups can be defined both when the data are read in the form of a presence/absence matrix, or as point records for each species; in either case, the string *groups* and a list of groups (enclosing within braces the list of species numbers that belong to each group) must follow after the data. For N species, as many as N–2 groups can be defined (i.e. this amounts to a fully resolved phylogenetic tree for the N species). In the format for point records (*.xyd files), the species are named as their records are defined. When the data are given in the form of a presence/absence matrix, the names can be contained in a separate file, with the number for each species to be named followed by its name (within square brackets); to input this file (e.g. *namelist.txt*) to VNDM, use *–names namelist.txt* as last argument to the program. When the species are named and higher groups are defined, then higher groups are automatically named with the string of species numbers that belong to the group.

A real example

Data set

To illustrate the approach just described, we have used a data set for North and Central American

mammals, taken from Escalante et al. (2010). The data set consists of 744 species (the species-only dataset of Escalante et al., 2010), with the generic nomenclature changed to be up to date. A total of 97 nodes representing groups taken from the literature (see below) were added to the 744 species. The grid used is of 4° × 4°, with only presence/absence data (i.e. no "assumed" records). It includes also some species present in South America and Europe. Escalante et al. (2010) included some higher taxa as separate units in alternative data sets; this is the best they could do with the implementation then available, but it is subject to the problem of redundancy discussed above. The full data set and group definitions can be downloaded (as Supplementary Material) from http://www.lillo.org.ar/phylogeny/published/Higher_taxa.zip.

Phylogenies

The higher groups used in this example were taken from Fumagalli et al. (1999), Alexander and Riddle (2005), Rezaei (2007), Fabre et al. (2012) and Melo-Ferreira et al. (2014). Note that, for a group to be included in the analysis, all the species of the group (with their full distributions) should be included in the distributional data set. The present analysis therefore did not include any tree-nodes for which some of the species were absent from the distributional data set. Excluding from the distributional data set a species that belongs to a monophyletic group amounts to excluding part of the distribution of the group. The only exception is when the distribution of the species excluded overlaps completely with that of the species included. In the distributional data set, this was the case, for example, for the pocket-mouse genus *Chaetodipus*, three species of which (*C. dalquesti*, *C. eremicus*, and *C. rudinoris*) were not included in the distributional data set. However, because the distribution of those three species overlaps with those of the species already included, the distribution of the group (genus) *Chaetodipus* in the analysis is already complete. Another caveat is for the complementary problem, of cases where the phylogenetic analysis includes only some of the taxa present in the distributional data set. An example is in the genus *Lepus*; the tree of Melo-Ferreira et al. (2014) defined *L. americanus* + *L. californicus* as a monophyletic group, and this group was used for the present data set. However, the tree of Melo-Ferreira et al. did not include some of the species of *Lepus* in the present data set (*L. alleni*, *L. callotis*, *L. flavigularis*, and *L. insularis*), which are then effectively assumed in the present paper to *not* belong to the group of *L. americanus* + *L. californicus*. Thus, the definition of higher groups in any analysis intended for serious definition of areas (instead of intended to just exemplify a method, as the present analysis) should carefully consider the problem of phylogenies including some species absent from the distributional data set, and trees excluding some species present in the distributional data set.

Both reviewers of the present paper raised the question of the ages of origin of the taxa used in the analysis. As previously discussed by Szumik and Goloboff (2004: 968), areas of endemism have not been based traditionally on the notion of vicariance, and the causal factor producing the given distributional pattern may, but need not, be history or vicariance. The present type of analysis is merely intended to detect congruence in the distributions, and information on ages is not required for doing so. Of course, knowing the ages of the groups will be important when investigating the causes for the concordance in distributions, or in other types of biogeographical study (e.g. analyes of vicariance).

Analysis

The parameters used in the analysis are as follows: areas must have an endemicity score of 2 or more, with two or more endemic species; counting taxa as endemic only if individual score is 0.4 or higher; not using edge proportions; keeping overlapping subsets when 40% of taxa are unique; ten replications (with initial random seed = 1).

Results

The analysis produced a total of 232 distinct areas, of scores 2.1833 (with four endemic species) to 61.4472 (with 81 endemic species). Figure 3 shows one of the areas found (number 50 in the Supplementary Material), which roughly corresponds to the Madrean region (Takhtajan, 1986), with nine taxa recognized as endemic (VNDM also reports *Nyctinomops macrotis* as endemic, but the data set does not include its complete distribution, which further extends to southern South America; the species is thus not relevant for the analysis; it was left in the analysis to match the data set and species numbering in Escalante et al., 2010). Of the nine taxa, three are higher groups, consisting of genera (*Chaetodipus*, *Xerospermophilus*) or groups of species within a genus (*Perognathus*). One of those higher taxa is a group of five species of *Perognathus* identified as monophyletic by Alexander and Riddle (2005). Another group is the four species in the genus *Xerospermophilus*, identified as monophyletic by Fabre et al. (2012; note that Fabre et al. followed the older nomenclature, including these species in *Spermophilus*; the generic delimitation we follow here is based on Helgen et al., 2009). The clearest example of the advantage of using higher taxa is in the genus *Chaetodipus*;

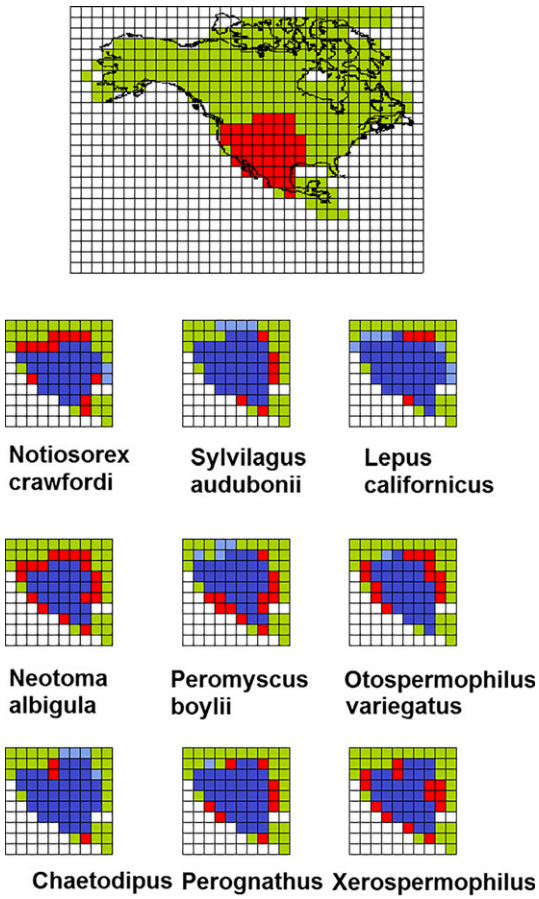

Fig. 3. List and distributions for species defining one of the areas of endemism ("Madrean region", found by the analysis of the data set of Escalante et al. (2010), with higher groups added). Of the nine endemic taxa, three are genera or higher groups (see Fig. 4 for details).

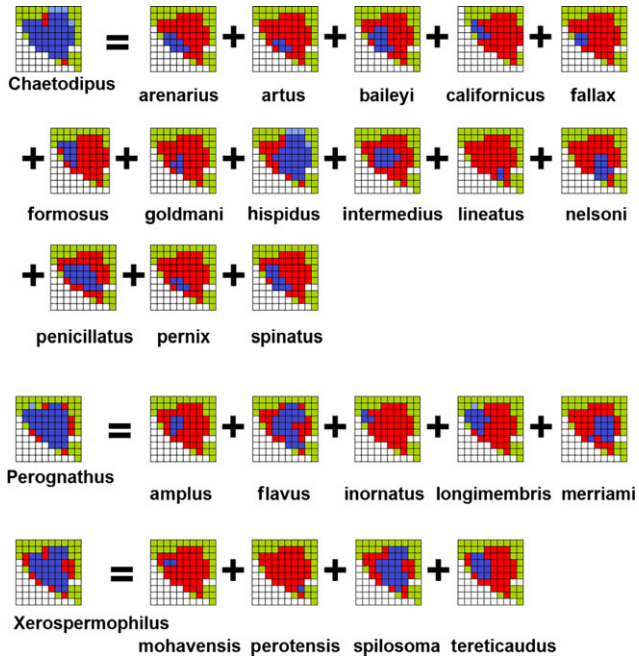

Fig. 4. Individual species composing the higher groups that determine the area of endemism shown in Fig. 3. Note that no single species has a distribution congruent with the area, but the overlapped distributions match the area well.

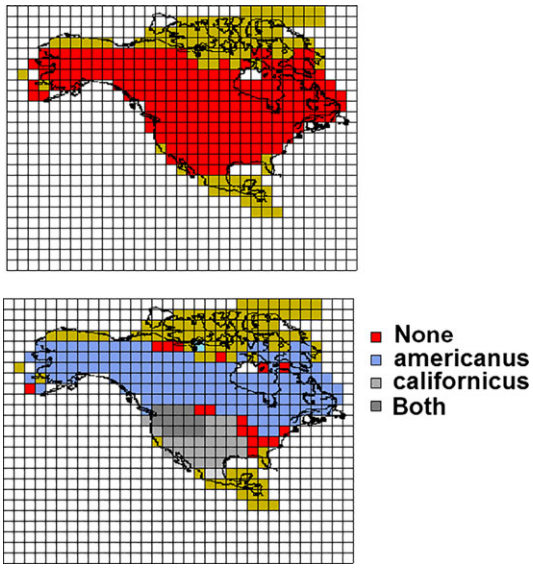

Fig. 5. An area of endemism (top map, "Taiga + Northern Forest") is almost perfectly congruent with the distribution of a monophyletic group (bottom map) formed by two species of *Lepus* (*L. americanus* and *L. californicus*).

none of the individual 14 species included in our analysis can be considered as endemic to the area (Fig. 4), but when all the distributions are overlapped, they fit the area reasonably well (with an endemicity score of 0.8100). The same happens with *Xerospermophilus*, and the monophyletic group within *Perognathus* (Fig. 4).

Another interesting case (Fig. 5) is a large area that extends across most of North America (a combination of the Taiga and Northern Forest ecological regions; number 199 in the Supplementary Material). Two species, *Castor canadensis* and *Ondathra zibethicus*, have congruent distributions delimiting this area. In addition to these, the group formed by *Lepus americanus* + *L. californicus* (but none of the individual two species of *Lepus*; Fig. 5, bottom) has a congruent distribution. This illustrates one of the points made in the previous sections: that large areas will often require considering higher taxa for a more precise delimitation.

When all the cases of higher taxa delimiting areas of endemism are considered, a total of 38 nodes (one-third of all the groups included) represent taxa endemic to some of the areas found, thus clearly helping to delimit the areas of endemism.

Acknowledgements

We thank the Consejo Nacional de Investigaciones Científicas y Técnicas (PIP 0687 to P.A.G.) for financial support, and acknowledge a collaborative grant (Dimensions US-Biota-São Paulo "Assembly and evolution of the Amazon biota and its environment: an integrated approach"), supported by the US National Science Foundation, National Aeronautics and Space Administration, and the Fundacão de Amparo a Pesquisa do Estado de São Paulo to Joel Cracraft and Lucía Lohman (with participation of P.A.G. and C.A.S.). We also thank José Carlos Guerrero Antúnez for guidance on exporting data to GIS programs, and the two reviewers (Pierre Deleporte, and Anonymous) for their comments on the manuscript.

References

Aagesen, L., Szumik, C., Goloboff, P., 2013. Consensus in the search for Areas of Endemism. J. Biogeogr. 40, 2011–2016.

Alexander, L.F., Riddle, B.R., 2005. Phylogenetics of the new world rodent family Heteromyidae. J. Mammal. 86, 366–379.

Carine, M.A., Humphries, C.J., Guma, I.R., Reyes-Betancort, J.A., Santos Guerra, A., 2009. Areas and algorithms: evaluating numerical approaches for the delimitation of areas of endemism in the Canary Islands archipelago. J. Biogeogr. 36, 593–611.

Escalante, T., Rodríguez, G., Szumik, C., Morrone, J.J., Rivas, M., 2010. Delimitation of the Nearctic Region according to mammalian distributional patterns. J. Mammal. 91, 1381–1388.

Fabre, P.H., Hautier, L., Dimitrov, D., Douzery, E.J.P., 2012. A glimpse on the pattern of rodent diversification: a phylogenetic approach. BMC Evol. Biol. 12, 88.

Fumagalli, L., Taberlet, P., Stewart, D.T., Gielly, L., Hausser, J., Vogel, P., 1999. Molecular phylogeny and evolution of *Sorex* shrews (Soricidae: Insectivora) inferred from mitochondrial DNA sequence data. Mol. Phylogenet. Evol. 11, 222–235.

Goloboff, P., 2002. NDM and VNDM: programs for the identification of areas of endemism, vers. 1.6. Program and documentation, available at www.lillo.org.ar/phylogeny.

Guedes, T.B., Sawaya, R.J., Nogueira, C.C., 2014. Biogeography, vicariance and conservation of snakes of the neglected and endangered Caatinga region, north-eastern Brazil. J. Biogeogr. 41, 919–931.

Hausdorf, B., Hennig, C., 2003. Biotic element analysis in biogeography. Syst. Biol. 52, 717–723.

Helgen, K.M., Cole, F.R., Helgen, L.E., Wilson, D.E., 2009. Generic revision in the Holarctic ground squirrel genus *Spermophilus*. J. Mammal. 90, 270–305.

Melo-Ferreira, J., Vilela, J., Fonseca, M.M., da Fonseca, R.R., Boursot, P., Alves, P.C., 2014. The elusive nature of adaptive mitochondrial DNA evolution of an arctic lineage prone to frequent introgression. Genome Biol. Evol. 6, 886–896.

Morrone, J.J., 1994. On the identification of areas of endemism. Syst. Biol. 43, 438–441.

Navarro, F., Cuezzo, F., Goloboff, P., Szumik, C., de Grosso, M.L., Quintana, M., 2009. Can insect data be used to infer areas of endemism? An example from the Yungas of Argentina Rev. Chil. Hist. Nat. 82, 507–522.

Rezaei, H.R., 2007. Phylogénie moléculaire du Genre *Ovis* (mouton et mouflons), implications pour la conservation du genre et pour l'origine de l' espèce domestique. Thèse de doctorat en Biodiversité, écologie, environnement.

Szumik, C., Goloboff, P., 2004. Areas of endemism: improved optimality criteria. Syst. Biol. 53, 968–977.

Szumik, C., Cuezzo, F., Goloboff, P., Chalup, A., 2002. An optimality criterion to determine areas of endemism. Syst. Biol. 51, 806–816.

Szumik, C., Aagesen, L., Casagranda, D., Arzamendia, V., Baldo, D., Claps, L., Cuezzo, F., Díaz Gómez, J.M., Di Giacomo, A., Giraudo, A., Goloboff, P., Gramajo, C., Kopuchian, C., Kretzschmar, S., Lizarralde, M., Molina, A., Mollerach, M., Navarro, F., Nomdedeu, S., Panizza, A., Pereyra, V., Sandoval, M., Scrocchi, G., Zuloaga, F., 2012. Detecting areas of endemism with a taxonomically diverse data set: plants, mammals, reptiles, amphibians, birds, and insects from Argentina. Cladistics 28, 317–329.

Takhtajan, A., 1986. Floristic Regions of the World (translated by T. J. Crovello & A. Cronquist). University of California Press, Berkeley.

Wallace, A.R., 1876. The Geographical Distribution of Animals with a Study of the Relations of Living and Extinct Faunas as Elucidating the Past Changes of the Earth's Surface. MacMillan and Co., London, Vol. 2.