Journal of Biogeography (J. Biogeogr.) (2013)

GUEST EDITORIAL

Consensus in the search for areas of endemism

Lone Aagesen[1]*, Claudia Szumik[2] and Pablo Goloboff[2]

[1]Instituto de Botánica Darwinion (CONICET-ANCEFN), San Isidro, B1642HYD, Buenos Aires, Argentina, [2]Instituto Superior de Entomología (INSUE-CONICET), 4000, Tucumán, Argentina

*Correspondence: Lone Aagesen, Instituto de Botánica Darwinion (CONICET-ANCEFN), Labarden 200, CC22, San Isidro, B1642HYD Buenos Aires, Argentina.
E-mail: laagesen@darwin.edu.ar

ABSTRACT

For ambiguous data sets, methods to determine areas of endemism based on an optimality criterion may result in large numbers of candidate areas, and thus some kind of consensus technique is required to summarize those results. This paper presents a formal description of two possible algorithms or rules for area consensus, which merge candidate areas if they share a user-defined percentage of the species that define each candidate area. The two consensus rules summarize ambiguity in different ways. Applying the 'tight' rule will result in consensus areas defined by species present in nearly all cells, but in cases where there is significant conflict the result may be a high number of distinct consensus areas. The 'loose' consensus rule is more agglomerative and will result in fewer consensus areas, combining areas when overlapping distribution patterns exist. Depending on the aim and scale of the analysis, the two consensus rules can be used either to delimit areas of endemism with sharp boundaries or to identify diffuse and gradually replacing biogeographical patterns. These two different approaches are discussed and demonstrated using real data.

Keywords

Areas of endemism, biogeography, conflicting species distributions, consensus algorithms, consensus rules, distributional congruence, VNDM.

INTRODUCTION

In the last few decades, a number of formal methods have been proposed to identify areas of endemism (Morrone, 1994; Crisp *et al.*, 2001; Linder, 2001; Szumik *et al.*, 2002; Hausdorf & Hennig, 2003) or distribution patterns akin to endemism (Dos Santos *et al.*, 2008). There are several interesting debates on areas of endemism, including their ontological nature (Crother & Murray, 2011, 2013). Defining areas of endemism is considered fundamental for historical and ecological biogeography (Crisp *et al.*, 2001; Crisci *et al.*, 2003), but an additional urge to define these areas comes from conservation studies, especially in cases where the protected area systems are focusing on diversity hotspots and endemism is found outside the highly diverse regions (Orme *et al.*, 2005; Whittaker *et al.*, 2005; Swenson *et al.*, 2012).

The various methods aimed at defining areas of endemism, differing both in details and in their general approach, have been compared and discussed elsewhere (e.g. Linder, 2001; Carine *et al.*, 2008; Escalante *et al.*, 2009; Casagranda *et al.*, 2012). However, one often overlooked but important issue is that real data sets at any geographical and taxonomic scale include ambiguous or contradicting patterns, as revealed by distribution analyses if applying an optimality criterion to search for multiple equal optimal solutions (García-Barros *et al.*, 2002; Rovito *et al.*, 2004; Carine *et al.*, 2008; Aagesen *et al.*, 2009; Szumik *et al.*, 2012).

Parsimony analysis of endemicity (PAE; Morrone, 1994) and the optimality criterion implemented in the open-source program VNDM by Szumik *et al.* (2002) are currently the only methods that explicitly explore and present ambiguity in distribution data when searching for areas of endemism. As PAE uses algorithms from phylogenetics to search for areas of endemism, it deals with multiple optimal solutions by the use of well-known consensus techniques from that conceptual framework. However, optimality criteria developed within the context of phylogenetic analysis carry several conceptual and practical problems when applied to the definition of areas of endemism (Szumik *et al.*, 2002). To avoid these problems, VNDM (Szumik *et al.*, 2002; Szumik & Goloboff, 2004) focuses on using an optimality criterion explicitly developed for evaluating candidate areas of endemism. The use of an optimality criterion implies that multiple candidate areas will be found in the case of data ambiguity. Authors have dealt differently with multiple solutions, either discussing a large number of areas or applying

doi:10.1111/jbi.12172

Journal of Biogeography

consensus techniques that have been recently made available in VNDM but not formally described (e.g. Domínguez *et al.*, 2006; Carine *et al.*, 2008; Ferrari *et al.*, 2010).

The aim of this paper is to discuss the conceptual background that consensus techniques must consider when applied to the context of species distributions, as well as providing a formal description of the details and implications of the algorithms available in VNDM. We use real data to illustrate two different consensus techniques and rules to be used depending on the aim and scale of the study.

VNDM

The method implemented in VNDM (Goloboff, 2004) is grid-based, and searches for 'areas' (= sets of cells) that are congruent with the distribution of as many species as possible. VNDM uses as input a list of species that includes georeferenced locations for each species. The georeferenced locations are transformed automatically into presence/absence in cells of the grid. To evaluate a candidate area, VNDM assigns a score to each species, depending on how well the species fits the area, with absences in part of the area as well as presence in cells outside the area penalized; the strength of penalization is user defined. Any area can receive an endemicity score, *E*, which is the sum of the scores of the supporting species. The value of *E* therefore improves both with the number of species concordant with the area, as well as with the degree of concordance between the area and those species.

One implication of using an optimality criterion is that areas that differ only in the presence or absence of a few cells (and thus are almost identical) may have the same value of *E*. This (as noted by Casagranda *et al.*, 2012) is an inescapable consequence of using an optimality criterion in any combinatorially intensive problem, with a perfect parallel in the multiple optimal trees often found in phylogenetic analyses (summarized by means of consensus trees). Therefore, when the output is several hundreds of possible areas, consensus techniques are required to summarize the results.

Consensus rules and cut-off values

The 'consensus' may focus on different aspects to be summarized. In the present case, it is desirable to establish some form of 'identity' of the areas to be placed together in a 'consensus area'. Doing so by similarity in cell composition alone would have the problem that, depending on the underlying species distributions, minor but consistent differences among sets of candidate areas could represent truly different areas of endemism. Thus, it seems preferable to operate at the level of supporting species, combining sets of candidate areas into consensus areas if they share some proportion of their respective defining species. Because under the criterion of VNDM (Szumik & Goloboff, 2004) a species will be considered as 'endemic' of an area only if it has a good spatial concordance with the area, then using the species takes into account, in an indirect but definite way, the spatial similarity of the areas to be included in one consensus. In the approach implemented in VNDM, the user defines the proportion of species that areas must share to merge them in a consensus area. As this proportion approaches unity, areas have to be more similar to be merged together. For proportions of shared species below unity, two possibilities arise.

Tight consensus rule

The strictest rule only adds a candidate area to the set of areas to be merged in a consensus if the area shares the selected percentages of defining species with *each* of the other areas in the consensus. As an example, Fig. 1a shows six individual areas supported by a total of 11 species distributed as seen in Table 1. Areas 2 and 3 share 75% of their species and will group into a consensus area using cut-off values equal to or lower than 75%. No other candidate area shares 75% of their defining species; therefore, under this cut-off, VNDM will display the consensus of areas 2 + 3, plus the remaining four candidate areas as individual areas (Fig. 1b). Note that the two candidate areas must share 75% of the union of their defining species, hence areas 5 and 6 do not group into a consensus under the 75% cut-off value: five species define these two areas, of which three are shared (Table 1), and they will only group under a cut-off value of 60% or lower. However, candidate area 1 also shares 60% of its defining species with candidate area 5 (three out of five species shared) and 60% of its defining species with candidate area 6 (also three out of five species shared; note these are not the same three species shared with area 5). Therefore, under a cut-off value of 60% the areas 1, 5 and 6 are merged into a single consensus area (Fig. 1c). VNDM will furthermore display consensus area 2 + 3 as well as candidate area 4 that shares less than 60% of its defining species with the other candidate areas. Candidate area 4 is more species-rich than the remaining areas but combines with area 6 under the cut-off value of 50% (sharing with area 6 three out of six species). However, area 6 still shares 50% (or more) of its defining species with areas 1 and 5; therefore, under this cut-off value, VNDM will display the distribution overlap in area 6 by presenting both consensus 1 + 5 + 6 and consensus 4 + 6 (as well as consensus area 2 + 3). As the candidate areas 1, 4, 5 and 6 share some species they will eventually group in a consensus area under sufficiently low cut-off values (28% or less; Fig. 1e). On the contrary, candidate areas 2 and 3 do not share any species with candidate areas 4, 5 or 6; hence, these areas will never group under the tight rule. However, as both candidate areas 2 and 3 share a single species with area 1, they will group with area 1 under the lowest possible cut-off value (14%, Fig. 1f), the cut-off under which consensus area 1 + 4 + 5 + 6 is still displayed.

Loose consensus rule

The second, more agglomerative consensus rule implemented in VNDM considers a candidate as part of a consensus set if it shares the selected percentages of defining species with at

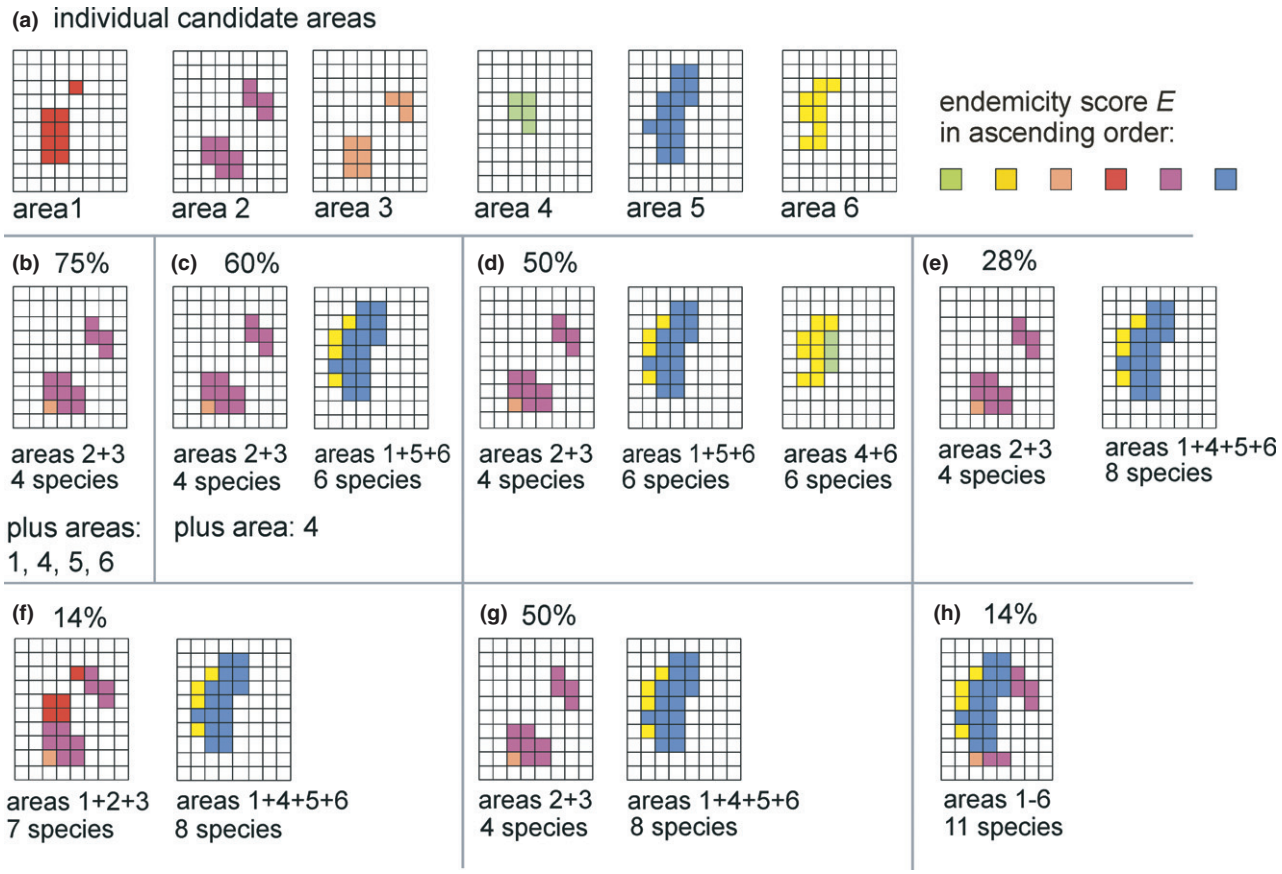

Figure 1 Hypothetical example to illustrate the result of applying different consensus rules and cut-off values to the 11 species shown in Table 1. (a) Six candidate areas each supported by the species in Table 1. (b–f) Output using the 'tight consensus rule' under the cut-off values of (b) 75%, (c) 60%, (d) 50%, (e) 28% and (f) 14%. (g–h) Output using the 'loose consensus rule' under the cut-off values (g) 50% and (h) 14%.

Table 1 Hypothetical example showing the distribution of 11 taxa in six different candidate areas. The effects of applying different consensus rules and cut-off values to these data are illustrated in Fig. 1.

	Candidate area					
	1	2	3	4	5	6
Taxon 1		x				
Taxon 2	x			x	x	x
Taxon 3				x		
Taxon 4				x		x
Taxon 5		x	x			
Taxon 6	x			x	x	x
Taxon 7					x	
Taxon 8	x				x	x
Taxon 9				x		
Taxon 10	x	x	x			
Taxon 11		x	x			

least *one other* set in the consensus (Fig. 1g,h). In the example, the two consensus rules will produce the same results under cut-off values above 50%, where no conflicting patterns were detected among the candidate areas. Using a 50% cut-off value, under which candidate area 6 shares some species with both area 1 and 5 and some with area 4, the loose rule merges all four candidate areas into the same consensus, as opposed to the tight rule, which displays two consensus areas (compare Fig. 1d and 1g). Likewise, under the lowest possible cut-off values (sharing a single species) the loose rule merges all candidate areas into a single consensus area, in contrast with the tight rule, which displays the overlapping pattern in candidate area 1 by displaying two distinct consensus areas (compare Fig. 1f and 1h). Thus, the loose rule agglomerates more candidate areas into consensus sets, producing fewer consensus areas.

The alternative positions of candidate area 1 are caused by taxon 10 which is found in the candidate areas 1, 2 and 3. The final result is either two possible tight-consensus areas (1 + 2 + 3 and 1 + 4 + 5 + 6) or a single loose-consensus area including all candidate areas. However, suppose that taxon 10 scores highly in candidate areas 2 and 3 while fitting candidate area 1 poorly. In that case, simply ignoring the presence of taxon 10 in candidate area 1 would remove the conflict and none of the consensus rules would merge

the pattern in consensus area 2 + 3 with the pattern in consensus area 4 + 5 + 6.

By default, VNDM finds all area sets with two or more endemic species, following Platnick's (1991) definition of area of endemism. Therefore, one way to prevent the high numbers of tight-consensus areas generated by partly overlapping but poorly supported area sets would be searching for candidate areas with more supporting species while simultaneously rejecting poorly fitting species. If a minimum of four endemic species with a relatively high fit were required for the data in Fig. 1, candidate areas 1 and 3 would not be reported and the two consensus rules would settle on the same results either under cut-off value 28% (Fig. 1e) or cut-off value 50% (Fig. 1g).

Colour scale of fit

When consensus areas are requested, VNDM calculates the consensus areas and displays them. When doing so, VNDM keeps track of the scores of the individual candidate areas, and displays the original score of each area set through a colour scale. The colour code associated with each score-range is shown next to the consensus area (colours can be changed by the user with a double click on the box displaying each colour). Figure 1 shows schematically how each cell in a consensus area is coloured. When a cell is part of several areas combined into a consensus area, the colour of the cell in the consensus corresponds to the set of highest score.

VNDM rescales the colours for each consensus in order to display maximum possible details within each consensus area. The colour codes of two different consensus areas found under the same consensus rule and cut-off value are therefore not necessarily comparable – they are scaled and partitioned differently. The absolute score of each cell will, however, still be comparable between consensus areas even if these are derived from different consensus rules and/or cut-off values, as long as the consensus areas are calculated from the same pool of candidate areas. In the example of Fig. 2, described below, we have edited the colour codes so that each colour represents a single score range, to facilitate comparisons between consensus areas.

Relationships between consensus rules, species overlap and diffuse borders in a real data set

Which of the consensus rules should be presented and discussed depends on the aim and scale of the analysis. Ideally, the tight consensus rule should be used for identifying well-defined areas of endemism, as this ensures that at least some species are shared by all the individual candidate areas, thus ensuring greater consistency among the areas merged into a consensus. Although the species defining the loose consensus may well be found in only a small part of the consensus area resulting from the merger, the loose consensus is a useful tool to outline gradual species overlap and replacement among candidate areas (as is evident from Fig. 1). Areas of endemism defined by the tight rule may be feasible only in small-scale or regional studies, where the main aim is to identify and separate the core areas of endemism for further examination, as in Fig. 2. Loose consensus areas may be sufficiently detailed in large-scale studies, such as areas of endemism in the Central Andes or the Southern Cone. The alternative in such large-scale studies would be presenting several tight consensus areas or increasing the cell size at the expense of missing details in the distribution patterns.

Figure 2 illustrates how the consensus rules can be applied to explore conflicting distributions in real data using as an example a data set of vascular plants endemic to the southern Central Andes (data set from Aagesen *et al.*, 2012). The data set includes 540 species, but conflicting distributions produce 695 candidate areas under a cell size of half a degree, which group into 82 tight consensus areas with a cut-off value of 5%. Most areas are supported by about 10 species; thus, the overlap is not caused by a few weakly supported candidate areas.

Under the loose consensus rule with a cut-off of 5%, 663 of the area sets combine into the consensus shown in Fig. 2a. Despite the complexity of the region, the loose consensus clearly shows three main areas of endemism, with endemicity much higher than in the surrounding cells. The northern and central areas have high endemicity, with their cells marked with highest fit values, while the third southernmost area is coloured, indicating a much lower degree of endemicity.

The two areas of high endemicity are not sharply delimited – they are instead surrounded by diffuse borders, evident from the gradual decrease in endemism in the neighbouring cells, caused by an overlap in species distribution between the core area and surrounding cells. In contrast, the southern area is only surrounded by cells of low fit, showing an abrupt decrease in endemicity as the borders of the area are crossed. By using different cut-off values it is possible to isolate the three areas of high endemicity for further examination. Three different cut-off values are required in this case, owing to the variable degree of species overlap between the core and the surrounding cells in these three areas.

The smaller southern area of endemism has well-defined borders and is separated from the main consensus area with a cut-off of 25% (Fig. 2b,c). Once it is isolated from the main area with this cut-off, it can be seen that the southern area corresponds to two candidate areas with 19 defining species. At this cut-off, the main consensus area continues, including the two areas of high endemicity, but a northernmost consensus area (merging 19 candidate areas) now separates from the main consensus as a large area of low endemicity (defined by 22 species; Fig. 2d). The area in Fig. 2d partly overlaps with the northern area and was therefore not detectable in Fig. 2a.

Raising the cut-off value to 40%, the northern consensus area (including its diffuse borders), separates from the main consensus area, while the central area continues to be included in the main area (Fig. 2e,f). The cores of both the

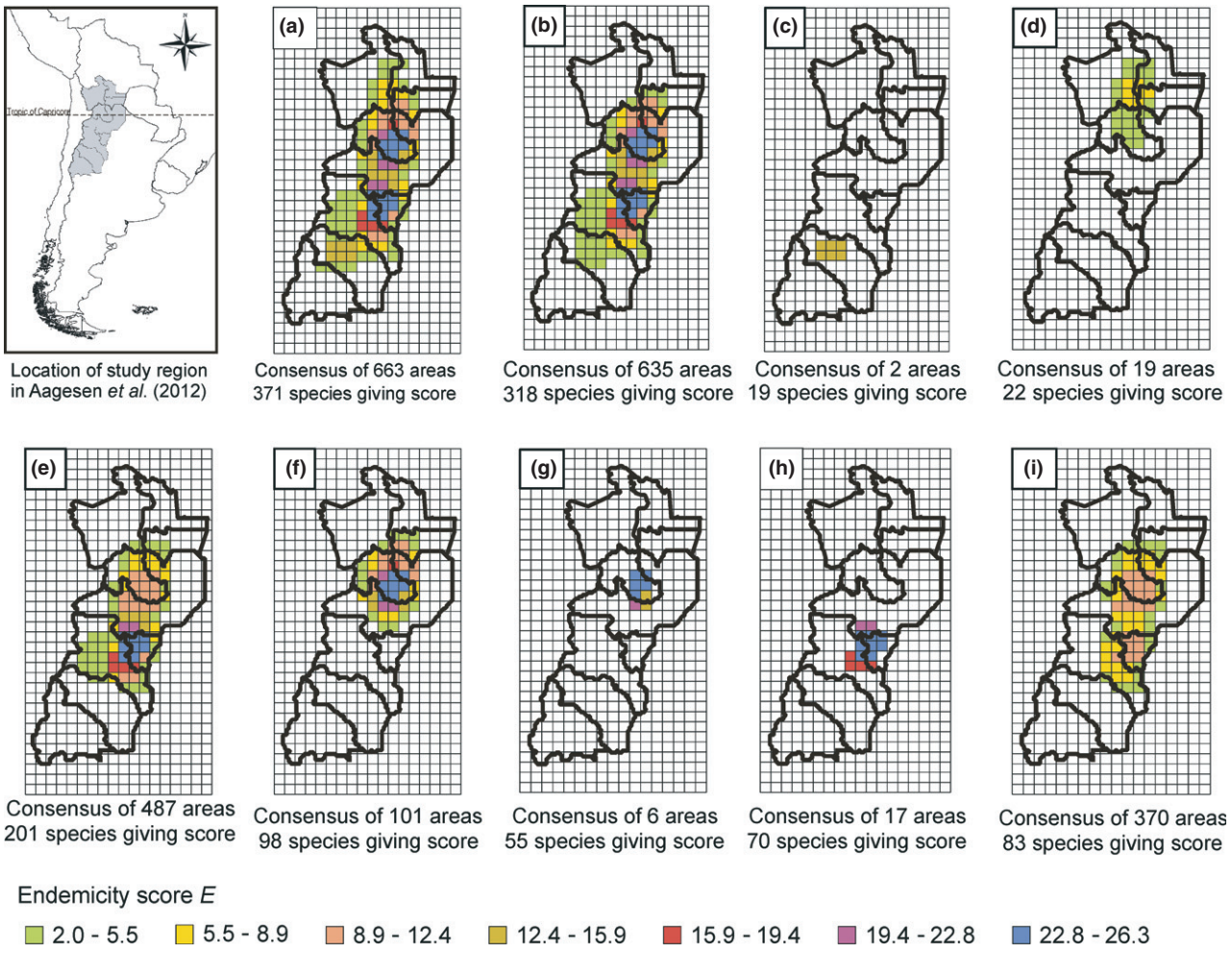

Figure 2 Consensus rules and cut-off values applied to distribution data of 540 vascular plant species from southern Bolivia and north-western Argentina. Different consensus areas obtained under the 'loose consensus rule' in a real data set under the cut-off values of (a) 5%, (b–d) 25%, (e,f) 40% and (g–i) 45%. The data set is from Aagesen *et al.* (2012) and available at http://www.zmuc.dk/public/phylogeny/TNT/More.

northern and central consensus areas separate only when the cut-off rises to 45% (Fig. 2g,h). Most of the remaining candidate areas are combined in a large consensus area of low endemicity (Fig. 2i). Note that although the area in Fig. 2i has low endemicity it is supported by a large number of species, indicating that many species fit the area but do so poorly. The number of candidate areas included in the consensus area of Fig. 2i (370 candidate areas), as well as in the core northern area of Fig. 2f (101 candidate areas), makes it evident that the ambiguity of the data set is mainly caused by the difficulty in delimiting these two distribution patterns.

CONCLUSIONS

The examples and discussion show that merging candidate areas of endemism into 'consensus' sets and displaying the overlapping sets with colour codes to indicate values of endemicity, provides a useful way to summarize the results of endemicity analyses that result in large numbers of candidate areas. The two consensus rules presented can be used depending on whether the main interest is in delimiting well-defined areas of endemism, or in identifying diffuse and gradually replacing biogeographical patterns.

ACKNOWLEDGEMENTS

We appreciated the help of the handling editor, Mark Carine, in processing and improving the manuscript, as well as the comments of three anonymous referees. We also thank the financial support of the Consejo Nacional de Investigaciones Científicas y Técnicas (PIP 0019 to C.S., PIP 0687 to Mercedes Lizarralde with participation of P.G., PIP 0207 to L.Aa.), the Agencia Nacional de Promoción Científica y Tecnológica (PICT 1314 to P.G.), and 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 Fundação de Amparo à Pesquisa do Estado de São Paulo to Joel Cracraft and Lucía Lohman, with participation of C.S. and P.G.

REFERENCES

Aagesen, L., Szumik, C., Zuloaga, F.O. & Morrone, O. (2009) Quantitative biogeography in the South America highlands—recognizing the Altoandina, Puna and Prepuna through the study of Poaceae. *Cladistics*, **25**, 295–310.

Aagesen, L., Bena, M.J., Nomdedeu, S., Panizza, A., López, R.P. & Zuloaga, F.O. (2012) Areas of endemism in the southern central Andes. *Darwiniana*, **50**, 218–251.

Carine, M.A., Humphries, C.J., Guma, I.R., Reyes-Betancort, J.A. & Santos Guerra, A. (2008) Areas and algorithms: evaluating numerical approaches for the delimitation of areas of endemism in the Canary Islands archipelago. *Journal of Biogeography*, **36**, 593–611.

Casagranda, M.D., Taher, L. & Szumik, C. (2012) Endemicity analysis, parsimony and biotic elements: a formal comparison using hypothetical distributions. *Cladistics*, **28**, 645–654.

Crisci, J.V., Katinas, L. & Posadas, P. (2003) *Historical biogeography: an introduction*. Harvard University Press, Cambridge, MA.

Crisp, M.D., Laffan, S., Linder, H.P. & Monro, A. (2001) Endemism in the Australian flora. *Journal of Biogeography*, **28**, 183–198.

Crother, B.I. & Murray, C.M. (2011) Ontology of areas of endemism. *Journal of Biogeography*, **38**, 1009–1015.

Crother, B.I. & Murray, C.M. (2013) Parsimony analysis of endemism under the "areas of endemism as individuals" thesis. *Cladistics*, doi:10.1111/cla.12023.

Domínguez, M.C., Roig-Juñent, S., Tassin, J.J., Ocampo, F.C. & Flores, G.E. (2006) Areas of endemism of the Patagonian steppe: an approach based on insect distributional patterns using endemicity analysis. *Journal of Biogeography*, **33**, 1527–1537.

Dos Santos, D.A., Fernández, H.R., Cuezzo, M.G. & Domínguez, E. (2008) Sympatry inference and network analysis in biogeography. *Systematic Biology*, **57**, 432–448.

Escalante, T., Szumik, C. & Morrone, J.J. (2009) Areas of endemism of Mexican mammals: reanalysis applying the optimality criterion. *Biological Journal of the Linnean Society*, **98**, 468–478.

Ferrari, A., Paladini, A., Schwertner, C.F. & Grazia, J. (2010) Endemism analysis of Neotropical Pentatomidae (Hemiptera, Heteroptera). *Iheringia, Série Zoologia*, **100**, 449–462.

García-Barros, E., Gurrea, P., Luciáñez, M.J., Cano, J.M., Munguira, M.L., Moreno, J.C., Sainz, H., Sanz, M.J. & Simón, J.C. (2002) Parsimony analysis of endemicity and its application to animal and plant geographical distributions in the Ibero-Balearic region (western Mediterranean). *Journal of Biogeography*, **29**, 109–124.

Goloboff, P. (2004) *NDM/VNDM: programs for identification of areas of endemism*. Programs and documentation. Available at: http://zmuc.dk/public/phylogeny/endemism.

Hausdorf, B. & Hennig, C. (2003) Biotic element analysis in biogeography. *Systematic Biology*, **52**, 717–723.

Linder, P. (2001) On areas of endemism, with an example from the African Restionaceae. *Systematic Biology*, **50**, 892–912.

Morrone, J.J. (1994) On the identification of areas of endemism. *Systematic Biology*, **43**, 438–441.

Orme, C.D.L., Davies, R.G., Burgess, M., Eigenbrod, F., Pickup, N., Olson, V.A., Webster, A.J., Ding, T.-S., Rasmussen, P.C., Ridgely, R.S., Stattersfield, A.J., Bennett, P.M., Blackburn, T.M., Gaston, K.J. & Owens, I.P.F. (2005) Global hotspots of species richness are not congruent with endemism or threat. *Nature*, **436**, 1016–1019.

Platnick, N.I. (1991) On areas of endemism. *Australian Systematic Botany*, **4**, xi–xii.

Rovito, S.M., Arroyo, M.T.K. & Pliscoff, P. (2004) Distributional modelling and parsimony analysis of endemicity of Senecio in the Mediterranean-type climate area of Central Chile. *Journal of Biogeography*, **31**, 1623–1636.

Swenson, J.J., Young, B.E., Beck, S. *et al.* (2012) Plant and animal endemism in the eastern Andean slope: challenges to conservation. *BMC Ecology*, **12**, 1–18.

Szumik, C. & Goloboff, P. (2004) Areas of endemism: an improved optimality criterion. *Systematic Biology*, **53**, 968–977.

Szumik, C., Cuezzo, F., Goloboff, P. & Chalup, A. (2002) An optimality criterion to determine areas of endemism. *Systematic Biology*, **51**, 806–816.

Szumik, C., Aagesen, L., Casagranda, D. *et al.* (2012) Detecting areas of endemism with a taxonomically diverse data set: plants, mammals, reptiles, amphibians, birds, and insects from Argentina. *Cladistics*, **28**, 317–329.

Whittaker, R.J., Araújo, M.B., Jepson, P., Ladle, R.J., Watson, J.E. & Willis, K.J. (2005) Conservation Biogeography: assessment and prospect. *Diversity and Distributions*, **11**, 3–23.

BIOSKETCHES

Lone Aagesen obtained her PhD from the University of Copenhagen in 2002 and is now a researcher for CONICET at the Darwinion Botanical Institute, Buenos Aires. She is interested in plant systematics, phylogenetics and species distributions, with emphasis on range-restricted species endemic to the Southern Cone.

Claudia Szumik received her PhD from Universidad de Tucumán in 1997 and is now a researcher for CONICET. Her main research interests are in the systematics of Embioptera (a poorly known insect order), and methodological aspects of historical biogeography.

Pablo Goloboff received his PhD from Cornell University in 1994. He is now a researcher for CONICET. His research focuses on the methodology of cladistic analysis and historical biogeography. He is author or co-author of several computer programs for parsimony or biogeographical analysis (TNT, NONA, PIWE, VNDM).

Editor: Mark Carine