Oecologia (2010) 162:463–477
DOI 10.1007/s00442-009-1474-y

COMMUNITY ECOLOGY - ORIGINAL PAPER

The empirical Bayes approach as a tool to identify non-random species associations

Nicholas J. Gotelli , Werner Ulrich

Received: 17 December 2008 / Accepted: 18 September 2009 / Published online: 15 October 2009

Abstract A statistical challenge in community ecology is to identify segregated and aggregated pairs of species from a binary presence–absence matrix, which often contains hundreds or thousands of such potential pairs. A similar challenge is found in genomics and proteomics, where the expression of thousands of genes in microarrays must be statistically analyzed. Here we adapt the empirical Bayes method to identify statistically significant species pairs in a binary presence–absence matrix. We evaluated the performance of a simple confidence interval, a sequential Bonferroni test, and two tests based on the mean and the confidence interval of an empirical Bayes method. Observed patterns were compared to patterns generated from null model randomizations that preserved matrix row and column totals. We evaluated these four methods with random matrices and also with random matrices that had been seeded with an additional segregated or aggregated species pair. The Bayes methods and Bonferroni corrections reduced the frequency of false-positive tests (type I error) in random matrices, but did not always correctly identify the non-random pair in a seeded matrix (type II error). All of the methods were vulnerable to identifying spurious secondary associations in the seeded matrices. When applied to a set of 272 published presence–absence matrices, even the most conservative tests indicated a fourfold increase in the frequency of perfectly segregated "checkerboard" species pairs compared to the null expectation, and a greater predominance of segregated versus aggregated species pairs. The tests did not reveal a large number of significant species pairs in the Vanuatu bird matrix, but in the much smaller Galapagos bird matrix they correctly identified a concentration of segregated species pairs in the genus *Geospiza*. The Bayesian methods provide for increased selectivity in identifying non-random species pairs, but the analyses will be most powerful if investigators can use a priori biological criteria to identify potential sets of interacting species.

Keywords Biogeography · Null model · C score · Presence–absence matrix · Statistical test

Communicated by Wolf Mooij.

Electronic supplementary material The online version of this article (doi:10.1007/s00442-009-1474-y) contains supplementary material, which is available to authorized users.

N. J. Gotelli (&)
Department of Biology, University of Vermont,
Burlington, VT 05405, USA
e-mail: ngotelli@uvm.edu

W. Ulrich
Department of Animal Ecology, Nicolaus Copernicus University
in Toruń, Gagarina 9, 87-100 Torun, Poland
e-mail: ulrichw@uni.torun.pl

Introduction

A major research focus in community ecology and biogeography has been the identification of non-random species associations in binary presence–absence matrices (Simberloff and Connor 1979; Gotelli and Graves 1996; Sfenthourakis et al. 2006). In these matrices, each row represents a species or taxon, each column represents a site or sample, and the entries indicate the presence (1) or absence (0) of a species in a site (McCoy and Heck 1987). There are patterns in such matrices that can be summarized by a single univariate metric, such as the nestedness of the matrix

(Patterson and Atmar 1986), or the C score (Stone and Roberts 1990), a measure of average pairwise species segregation.

In null model analysis (Gotelli 2001), the observed matrix is randomized or reshuffled, and the metric is

algorithm (Gotelli 2000), in which species occurrences are randomized, but the row sums (=species incidences) and column sums (=species richness per site) of the observed matrix are preserved. In benchmark tests for community metrics of nestedness (Ulrich and Gotelli 2007a) and co-occurrence (Gotelli 2000), this algorithm correctly identifies random matrices and structured matrices with acceptable type I and type II error frequencies. However, the constraint of fixed row and column totals introduces associations between species and sites that might distort the number of species pairs that fall outside the 95% CL (Ulrich and Gotelli 2007a).

Sequential Bonferroni correction (Benjamini and Yekutieli criterion)

The Bonferroni correction is a simple metric to reduce the FDER by dividing the significance level a by the total number of tests *r*

scores). In bin C, 32 species pairs had scores between 0.50 and 0.55, whereas 19.5 4.0 (mean 95% CL) were expected (26 pairs marks the upper 95% CL). In bin D, 13 species pairs were observed with scores between 0.70 and 0.75, but only 3.1 1.9 pairs were expected (eight pairs

species generate many species pairs: with species numbers from only 15–45, the resulting distribution will have between 100 and 1,000 unique species pairs.

Third, the Bayes methods (and the standard confidence interval methods) identify which species pairs are non-random, but they do not specify whether the pattern is one of segregation or aggregation. To classify species pairs as aggregated or segregated, we compared the observed C score with the mean of the simulated C scores for a particular species pair. Segregated pairs are those for which the observed C score was greater than the average simulated C score, and aggregated pairs are those for which the observed C score was less than the average simulated C score. Naturally, the majority of the segregated species pairs occur in the bins that are close to 1.0, and the majority of the aggregated species pairs occur in bins that are close to 0.0. These pairs represent cases of very strong segregation (perfect or near perfect checkerboard distributions) or very strong aggregation (complete or nearly complete overlap). However, as seen in Fig. 3, there is also a collectmple do not specifys606.12‘9arly wep550.3(represe(do)-3)-309424.6(veryn)-237.8*[(o-586.T*25197.8358.)-3ge37.8*[(of37.8*tha

of the primary pair changed the C score of the entire matrix. These secondary pairs represent associations between one species in the original matrix and one of the two new species added to the seeded matrix. We also counted how often C score identified all pairs as b

2008). The online appendix provides a spreadsheet with the original data matrix (Ulrich and Zalewski 2006) and fully documented output from the Pairs analysis illustrated in Fig. 1.

Results

Benchmark random matrices

Between 3.72 and 4.40% of the original M_N and M_E matrices had significantly aggregated or segregated species pairs as judged by the 95% CL benchmark (CL criterion) of the fixed–fixed null model (Table 1). The BY criterion reduced this fraction to 0.97% for the M_E and 1.65% for the M_N

Table 3

Table 4 Percentage of species pairs ranging outside the 95% CLs of the fixed–fixed null model for each combination of matrix (M_E and M_N) and seeded species pairs (segregated and aggregated) that were added to a random matrix

	M_E50								M_N50							
	CL criterion		Bayes M criterion		Bayes CL criterion		BY criterion		CL criterion		Bayes M criterion		Bayes CL criterion		BY criterion	
	All pairs (%)	Seeded pairs (%)	All pairs (%)	Seeded pairs (%)	All pairs (%)	Seeded pairs (%)	All pairs (%)	Seeded pairs (%)	All pairs (%)	Seeded pairs (%)	All pairs (%)	Seeded pairs (%)	All pairs (%)	Seeded pairs (%)	All pairs (%)	Seeded pairs (%)

number of aggregated or segregated species pairs exceed the expected numbers (positive effect sizes). Hence for at least 218 matrices we did not observe a significant trend towards species segregation or aggregation when using a species pair approach. However, in 107 matrices we did find a significant matrix-wide C score (cf. Gotelli and McCabe 2002), even when no individual cases of strongly segregated species pairs could be detected.

Vanuatu and Galapagos matrix analyses

The Vanuatu bird matrix contains 56 species and 28 sites (Diamond and Marshall 1976). As demonstrated in other analyses (Stone and Roberts 1990; Gotelli and Entsminger 2001

Table 7

(24%). In contrast, five of the seven significant segregated pairs identified by the Bayes M criterion were congeners (71%) (v^2 contingency test: v^2 3.03, *P* 0.08). Sanderson (2000) and Sfenthourakis et al. (2006) reported similar results for this data set using the CL criterion, although Sanderson (2000) used a different null model algorithm.

Discussion

Pairwise tests of species co-occurrence patterns invariably reveal statistically significant associations in random matrices using the simple 95% CL criterion (Table 1). The sequential Bonferroni, Bayes M, and Bayes CL criteria substantially reduce such occurrences, although they do not entirely eliminate them from random matrices. However, these analyses reveal the unavoidable trade-off between type I and type II statistical errors. For random matrices that were seeded with a non-random pair of species, the simple CL criterion did the best job of recovering these patterns, whereas the Bonferroni and Bayes methods did not detect the non-random pair in a substantial number of cases.

One difficulty is that all four of the methods detected false "secondary pairs" of species associations that emerged when a single non-random association was added to the matrix (Table 1). This result probably reflects, in part, the complex non-independence among all species pairs when the null model preserves fixed row and column totals. However, these statistically significant secondary pairs were more of a problem for aggregated than segregated distributions. Previous authors have discussed the possibility of a "dilution effect" in null model analysis in which significantly segregated species pairs are not detected because too many pairwise comparisons are made between pairs of species that are not interacting (Diamond and Gilpin 1982; Colwell and Winkler 1984). However, our results suggest there may well be a "concentration effect" because the addition of a single non-random species pair to a random matrix may generate a number of significant secondary pairs.

The analysis of the 13ufi[5.7TD[selishedfi[5.8.2(matrice)-10.3(s)[5.7 concentration of both highly segregated and highly aggregated species pairs, but also a set of species pairs that showed weaker, but still highly, patterns of overlap (Fig. 3). However, most of the significant pairs were concentrated in a relatively small number of matrices (Table 6). In many cases, the overall C score of the matrix maybe highly significant evenfi[5.8.4(thoug)-6.6(hfi[5.2.5(few)-266.1 ostracods, and fish), whereas only two of the 15 most segregated matrices (13.3%) were poikilotherms (Table 6). These resultsare consistent with matrix-wide patterns of species segregationfi[551.2(()0(G)-7.4(otelli)-256.4(and)-26 plant matrices than for [oikilotherms (invertebrate, amphibian, reptile, and fish matrices). Although all matrices contain segregated, random, and aggregated species pairs, the frequencies of these pairwisens are consistent with the overall matrix score.

However, pairwise analyses may not always be concordant with overall matrix scores. Both the Galapagos and

Vanuatu bird matrices have significant matrix-wide segregation, but the pairwise analysis of the Vanuatu matrix revealed very few significant pairs of species, which are ecologically and phylogenetically heterogeneous. In contrast, the Bayes M criterion identified seven significant pairs in the much smaller Galapagos matrix (Table 7). Five of these seven species pairs were concentrated within the genus *Geospiza*, which is one of the few examples of a competitively structured community that has been supported by extensive null model analysis (Simberloff and Connor 1981; Schluter and Grant 1984; Sanderson 2000).

Although the Bayes criteria and the sequential Bonferroni test do a better job of guarding against type I errors than the simple CL criterion, all of the methods proposed here must be used with caution. Even the most stringent criteria still detected a small number of unusual pairs in a large random matrix, and random matrices that were seeded with significant species pairs generated spurious statistical associations with other species in the matrix.

Perhaps it is asking too much of a statistical analysis to reveal biologically meaningful pairwise associations with no other information than a binary presence–absence matrix. A similar limitation has emerged in regression analyses and model selection. Whereas ecologists often use stepwise criteria to select a subset of meaningful predictor variables, these methods do not always identify the correct underlying model. A more powerful approach is to specify a priori a set of potential biological models, fit them to the data, and then use model selection criteria to rank them or distinguish between them (Burnham and Anderson 2002; Shipley 2002). For presence–absence matrices, the best strategy might be to identify ahead of time guilds or subsets of potentially interacting species and restrict the analysis to these pairs.

Acknowledgments This work was supported in part by a grant from the Polish Ministry of Science to W. U. (KBN, 2 P04F 039 29). N. J. G. was supported by NSF grant 0541936. We thank Aaron Ellison for calling our attention to Efron (2005). The manuscript was improved by comments from two anonymous reviewers and associate editor W. M. Mooij.

References

Abbott I, Black R (1980) Changes in species composition of floras on islets near Perth, Western Australia. J Biogeogr 7:399–410

Atmar W, Patterson BD (1995) The nestedness temperature calculator: a visual basic program, including 294 presence absence matrices. AICS Research Incorporate and The Field Museum. http://www.aics-research.com/nestedness/tempcalc.html

Bacallado JJ (1976) Notas sobre la distribucion y evolucion de la avifauna Canaria. In: Kunkel G (ed) Biogeography and ecology in the Canary Islands. Junk, The Hague, pp 13–431

Beard JS (1948) The natural vegetation of the Windward and Leeward Islands. Oxford For Mem 21:1–192

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B 57:289–300

Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29:1165–1188

Brown JH, Kurzius MA (1987) Composition of desert rodent faunas: combinations of coexisting species. Ann Zool Fenn 24:227–237

Burnham KP, Anderson DR (2002) Model selection and inference: A practical information-theoretic approach. Springer, New York

Burns KC (2007) Patterns in the assembly of an island plant community. J Biogeogr 34:760–768

Cameron RAD (1992) Land snail faunas of the Napier and Oscar Ranges, Western Australia; diversity, distribution and speciation. Biol J Linn Soc 45:271–286

Colwell RK, Winkler DW (1984) A null model for null models in biogeography. In: Strong Jr, Simberloff D, Abele LG, Thistle AB (eds) Ecological communities: conceptual issues and the evidence. Princeton University Press, Princeton, pp 344–359

Connor EF, Simberloff D (1979) The assembly of species communities: chance or competition? Ecology 60:1132–1140

Crowe TM (1979) Lots of weeds. J Biogeogr 6:169–181

Descimon H (1986) Origins of Lepidopteran faunas in the high tropical Andes. In: Vuilleumier F, Monasterio M (eds) High altitude tropical biogeography. Oxford University Press, Oxford, pp 500–532

Diamond JM (1975) Assembly of species communities. In: Cody ML, Diamond JM (eds) Ecology and evolution of communities. Harvard University Press, Cambridge, pp 342–444

Diamond JM, Gilpin ME (1982) Examination of the "null" model of Connor and Simberloff for species co-occurrences on islands. Oecologia 52:64–74

Diamond JM, Marshall AG (1976) Origin of the New Hebridean avifauna. Emu 76:187–200

Efron B (2005) Bayesians, frequentists, and scientists. J Am Stat Assoc 100:1–5

Gotelli NJ (2000) Null model analysis of species co-occurrence patterns. Ecology 81:2606–2621

Gotelli NJ (2001) Research frontiers in null model analysis. Glob Ecol Biogeogr 10:337–343

Gotelli NJ, Ellison AE (2004) A primer of ecological statistics. Sinauer, Sunderland

Gotelli NJ, Entsminger GL (2001) Swap and fill algorithms in null model analysis: rethinking the knight's tour. Oecologia 129:281–291

Gotelli NJ, Entsminger GL (2003) Swap algorithms in null model analysis. Ecology 84:532–535

Gotelli NJ, Graves GR (1996) Null models In ecology. Smithsonian Institution Press, Washington

Gotelli NJ, McCabe DJ (2002) Species co-occurrence: a meta-analysis of J. M. Diamond's assembly rules model. Ecology 83:2091–2096

Haila Y, Järvinen O, Vaisanen RA (1980) Habitat distributions and species associations of land bird populations on the Aland Islands, SW Finland. Ann Zool Fenn 17:87–106

Hatt RT, Van Tyne J, Stuart LC, Pope CH, Grobman AB (1948) Island life: a study of the land vertebrates of the islands of eastern Lake Michigan. Cranbrook Institute of Science. Bloomfield Hills, MI

Higgins CL, Willig MR, Strauss RE (2006) The role of stochastic processes in producing nested patterns of species distributions. Oikos 114:159–167

Hocutt CH, Denoncourt RF, Stauffer JR (1978) Fishes of the Greenbriar River, West Virginia, with drainage history of the Central Appalachians. J Biogeogr 5:59–80

Kammenga JE, Herman MA, Ouborg NJ, Johnson L, Breitling R (2007) Microarray challenges in ecology. Trends Ecol Evol 22:273–279
Lehsten V, Harmand P (2006) Null models for species co-occurrence patterns: assessing bias and minimum iteration number for the sequential swap. Ecography 29:786–792
Manly BFJ (1991) Randomization and Monte Carlo methods in biology. Chapman and Hall, London
Manly BFJ (1995) A note on the analysis of species co-occurrences. Ecology 76:1109–1115
May RM (1975) Patterns of species abundance and diversity. In: Cody ML, Diamond JM (eds) Ecology and evolution of communities. Harvard University Press, Cambridge, pp 81–120
McCoy ED, Heck KL (1987) Some observations on the use of taxonomic similarity in large-scale biogeography. J Biogeogr 14:79–87
Miklós I, Podani J (2004) Randomization of presence–absence matrices: comments and new algorithms. Ecology 85:86–92
Moran MD (2003) Arguments for rejecting the sequential Bonferroni in ecological studies. Oikos 100:403–405
Murphy RW (1983) The reptiles: origins and evolution. In: Case TJ, Cody ML (eds) Island biogeography in the Sea of Cortez. University of California Press, Berkeley, pp 130–158
Patterson BD (1987) The principle of nested subsets and its implications for biological conservation. Conserv Biol 1:323–334
Patterson BD, Atmar W (1986) Nested subsets and the structure of insular mammalian faunas and archipelagos. In: Heaney LR, Patterson BD (eds) Island biogeography of mammals. Academic Press, London, pp 65–82
Patterson BD, Pacheco V, Solari S (1996) Distributions of bats along an elevational gradient in the Andes of south-eastern Peru. J Zool 240:637–658
Sanderson JG (2000) Testing ecological patterns. Am Sci 88:332–339
Sanderson JG (2004) Null model analysis of communities on