Coding and development issues in FALCON

Stephen J. Beckett*, Chris A. Boulton, Hywel T. P. Williams

July 11, 2014

College of Life and Environmental Sciences, University of Exeter, Exeter, UK

*author for correspondence: S.J.Beckett@exeter.ac.uk

In the creation of the FALCON statistical framework, we came across many decisions that needed to be made in the process to find the nestedness of a matrix and its significance. Some of the key issues we came across are outlined below:

- Some users of nestedness measures elect to leave in some nodes from their data that do not connect to any other nodes (e.g. (Flores et al., 2011)), whilst this may make sense from a specific data perspective these nodes do not actually form part of the interaction network and we believe they should be removed from the analysis - else why not add further spurious nodes?

- Sometimes nestedness analysis, especially in a biogeographic setting is used to determine the effects of particular ecological gradients on the nestedness pattern in the matrix - and thus the input data that is used is sorted along some predetermined ecological gradients. However, in other settings, these gradients typically are not known and instead the key question instead is related to how nested the interaction network actually is - by sorting rows and columns for maximum nestedness. This latter definition is more broad and applicable to the statistical structure of nestedness to be found in matrices than the former, which is not computable for several nestedness measures. However, we allow both kinds of analysis to be performed by FALCON.

- The FF null model offers several challenges - the foremost of which is to ensure that null matrices are fairly generated from the distribution of null matrices that exist. Miklós and Podani, 2004 show that using a trial swap methodology rather than just the swaps that exists gives a better distribution, but requires a large number of trial swaps. They also show that generating these types of null matrices is a Markov chain problem and suggest two other methodologies for generating these matrices that may be faster than the trial swap algorithm. However these methods complete in stochastic time and we found that whilst they were faster than trial swaps on random matrices with a near homogeneous degree distribution, they were exponentially slower at generating null ensembles for matrices that had a heterogeneous degree distribution (results not shown) - most pertinently this includes those matrices which are nested. For these reasons we decided to use a trial swap method using 30,000 trial swaps (or rows multiplied by columns if this is larger) as a spinup to find the first null matrix and subsequent null matrices are found after 5,000 additional trial swaps using advice given in (Gotelli and Ulrich, 2011). This is a much more computationally demanding null model than the others used in FALCON. There may be a clever way to choose between different algorithms for different matrices to improve performance for this type of null model; or new or improved algorithms may be of interest to researchers interested in this null model. Some matrices have no available swaps - in these cases performing the trial swap algorithm is a hindrance, as such we check that at least one swap is possible before beginning the trial swap algorithm. If no swaps are available the same measure score as for the initial matrix is assigned to all members of the ensemble such that a p-value of p=1 will be assigned.

- In order to address the FF null model in this way we changed the code system such that instead of generation of null matrices and measurement operations being performed independently of one another in statistical ensembles as in a previous version of FALCON(Beckett and Williams, 2013), that measurements are taken from within the null model ensemble generation functions.

- After we uploaded the first iteration of FALCON Strona et al. (2014) published the so-called curveball algorithm for generating null matrices under the FF null model. This algorithm resolves the difficulties with the algorithms detailed above and is both quicker and more robust than the previous implementations. In light of this we replaced the trial-swap method with that of the curveball algorithm.

- Although the NTC(Atmar and Patterson, 1993) was the first way to quantify nestedness using both row and column information and received much popularity due to the software that accompanied it - the actual methodology given for its computation was lacking as the source code was hidden away in the application; a black box. This lead to several subsequent publications which attempted to lay down a more concise methodology (Guimarães and Guimarães, 2006; Rodríguez-Gironés and Santamaría, 2006; Ulrich and Gotelli, 2007; Oksanen et al., 2013). These procedures also had accompanying software, but similar to the nestedness temperature calculator the actual source code was also hidden away in some of these applications. As such it is hard not to treat these as statistical black boxes. `nestedtemp (Oksanen et al., 2013)` was one of just two open source methodologies that we were able to find for calculating nestedness in a way similar to the original NTC, though (Ulrich and Gotelli, 2007) uses a pair of linear isoclines, rather than the curve in the original method. We therefore decided to use `nestedtemp` as the basis of the NTC used in FALCON, especially as its use is also recommended by the R bipartite package (Dormann et al., 2008).

- As highlighted in the description the choice for an adaptive ensemble may be non-trivial. Previously our methodology used the convergence of the average nestedness score in two sampling groups as an indicator for whether the sampling groups were representative of the population. Here instead we have chosen to perform a Mann-Whitney U test between the two sampling groups to question whether they appear to have come from the same discrete distribution (at the 10% significance level) as an indicator of their representation of the population. We believe this test is better as the underlying sampling space of nestedness measures is discrete and is not sensitive to just one descriptor of a distribution. There may be other and better ways to tell when a population is representatively sampled, but here we merely sow the seeds of this idea.

- Another new feature to this version of FALCON is that multiple nestedness measures can be taken within the same null ensemble using the same null matrices. Thus for the adaptive solver's condition to be satisfied - the two sampling groups must satisfy the Mann-Whitney U test for each of the specified nestedness measures. This could prove problematic if there was an exceptionally large number of measures (such that the condition might always fail from one measure by chance alone) or a nestedness measure was chosen that was unable to satisfy this condition (i.e. continuous or stochastic) and is something for potential FALCON developers to be aware of.

References

Flores, C. O., Meyer, J. R., Valverde, S., Farr, L., and Weitz, J. S. (2011) Statistical structure of host–phage interactions. *Proceedings of the National Academy of Sciences 108*, E288–E297.

Miklós, I., and Podani, J. (2004) Randomization of presence-absence matrices: comments and new algorithms. *Ecology 85*, 86–92.

Gotelli, N. J., and Ulrich, W. (2011) Over-reporting bias in null model analysis: a response to Fayle and Manica (2010). *Ecological Modelling 222*, 1337–1339.

Beckett, S. J., and Williams, H. T. P. (2013) Coevolutionary diversification creates nested-modular structure in phage–bacteria interaction networks. *Interface Focus 3*, 20130033.

Strona, G., Nappo, D., Boccacci, F., Fattorini, S., and San-Miguel-Ayanz, J. (2014) A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nature communications 5*.

Atmar, W., and Patterson, B. D. (1993) The measure of order and disorder in the distribution of species in fragmented habitat. *Oecologia 96*, 373–382.

Guimarães, P. R., and Guimarães, P. (2006) Improving the analyses of nestedness for large sets of matrices. *Environmental Modelling & Software 21*, 1512–1513.

Rodríguez-Gironés, M. A., and Santamaría, L. (2006) A new algorithm to calculate the nestedness temperature of presence–absence matrices. *Journal of Biogeography 33*, 924–935.

Ulrich, W., and Gotelli, N. J. (2007) Null model analysis of species nestedness patterns. *Ecology 88*, 1824–1831.

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., and Wagner, H. vegan: Community Ecology Package. 2013; R package version 2.0-10.

Dormann, C. F., Gruber, B., and Fruend, J. (2008) Introducing the bipartite Package: Analysing Ecological Networks. *R News 8*, 8–11.