Short Communication

A Java program for non-parametric statistic comparison of community structure

WenJun Zhang

School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China; International Academy of Ecology and Environmental Sciences, Hong Kong

E-mail: zhwj@mail.sysu.edu.cn, wjzhang@iaees.org

Received 19 June 2011; *Accepted 23 July 2011*; *Published online 1 September 2011*

IAEES

Abstract

The Java algorithm to statistically compare structure difference of two communities was presented in this study. Euclidean distance, Manhattan distance, Pearson correlation, Point correlation, quadratic correlation and Jaccard coefficient were included in the algorithm. The algorithm was used to compare rice arthropod communities in Pearl River Delta, China, and the results showed that the family composition of arthropods for Guangzhou, Zhongshan, Zhuhai, and Dongguan are not significantly different.

Keywords community; structure; non-parametric test; comparison; arthropods; Java; program.

1 Introduction

The structure of community refers to species composition, population size, etc., which is formed by environment/climate conditions, and dynamic interspecific/intraspecific interactions (Damgaard, 2011; Lüi, 2011; Rai, 2011; Watts and Worner, 2011; Zhang, 2011; Zhang and Chen, 2011). Comparison of structure difference between two communities is always important. Non-parametric statistics may be used in the difference comparison (Clarke, 1993; Schoenly and Zhang, 1999). In this article a Java algorithm, based on previous studies, was presented to statistically compare between-community structure difference.

2 Algorithm

The following two algorithms are used to compare the comprehensive difference in structure composition (i.e., taxa and individual number, randomness of taxa, taxa placement, etc.) between two communities (Clark, 1993; Schoenly and Zhang, 1999).

Assume that there are s taxa in both community 1 and community 2. The number of samples is m in community 1 and is n in community 2, in total there are $ts=m+n$ samples in combined community. Given a_{ij}, the individual number of taxon i in sample j, $i=1,2,...,s$; $j=1,2,...,ts$. Calculate the distance (similarity) between sample i and j, $i=1,2,...,ts-1$; $j=i,...,ts$. The following distance (similarity) measures, Euclidean distance, Manhattan distance, Pearson correlation, Point correlation, quadratic correlation and Jaccard coefficient can be calculated:

$$d_{ij}=(\sum_{k=1}^{s}(a_{ki}-a_{kj})^2/s)^{1/2}$$

$$d_{ij}=\sum_{k=1}^{s}|a_{ki}-a_{kj}|/s$$

$$d_{ij}=\sum^{s}_{k=1}((a_{ki}-a_{ibar})(a_{kj}-a_{jbar}))/(\sum^{s}_{k=1}(a_{ki}-a_{ibar})^2\sum^{s}_{k=1}(a_{kj}-a_{jbar})^2)^{1/2}$$

$$d_{ij}=(ad-bc)/((a+b)(c+d)(a+c)(b+d))^{1/2}$$

$$d_{ij}=\sin((a+d-(b+c))/(a+b+c+d)*3.1415926/2)$$

$$d_{ij}=(b_i+b_j)/(c_i+c_j-e)$$

$$i=1,2,...,ts-1;\ j=i,...,ts$$

In the last three measures, both sample i and sample j take values 0 or 1, i, j=1,2,...,s. a is number of both sample i and sample j take value 0, b is number of sample i takes 0 and sample j takes 1, c is number of sample i takes 1 and sample j takes 0, and d is number of both sample i and sample j take value 1. b_i is the non zero number present in sample i but not in sample j, b_j is the non zero number present in sample j but not in sample i, c_i and c_j is the non zero number in sample i and sample j respectively, and e is non zero number shared by sample i and sample j.

Let $b_k=d_{ij}$, i=1,2,...,ts-1; j=i,...,ts; k=1,2,...,(ts*ts-ts)/2+ts-1. Rank b_k from small to large values, then re-ranked b_k and its ranking value g_k are thus given, k=1,2,...,(ts*ts-ts)/2+ts-1. For each of re-ranked b_k, k=1, 2,..., (ts*ts-ts)/2+ts -1, if its corresponding two samples belong to the same community, then let h_k=1, g_k=k, or else let d_k=1, f_k=k. Given the number of h_k=1, is k_p, the number of d_k=1, is r_p, the sum of g_k is s_p, the sum of f_k is c_p. Calculate r measure:

$$r=4*(c_p/r_p-s_p/k_p)/(ts*(ts-1)),$$

then let r_0=r, i.e., observed r value. Using Monte Carlo technique, randomly divide all of d_{ij}, i=1,2,..., ts-1; j=1,2,...,ts, into two communities with random number of samples in first community, the first community has m_1 samples and the second community has ts-m_1 samples. Let b_k= d'_{ij}, where d'_{ij} is d_{ij} after randomization, i=1,2,...,ts-1; j=i,...,ts; k=1,2,..., (ts*ts-ts)/2+ts-1.

Repeat the above procedures from which the r for this randomization can be calculated. For v randomizations, record the total number of $r\geq r_0$ as w, and expected and standard deviation of r can be derived also. Finally, calculate the p value:

$$p=(w+1)/(v+1).$$

If p is less than 0.05, or 0.01, then the difference of structure composition between community 1 and community 2 is statistically significant.

The algorithm is implemented as a Java program, CommStrucComp, based on JDK 1.1.8, in which several classes and an HTML file is included (http://www.iaees.org/publications/software/index.asp). In community 1 and community 2 data files, the first column is taxon ID number, and the first row is sample ID number.

3 Application

We obtained a set of data investigated in rice fields of four cities of Pearl River Delta, Guangzhou (23 samples), Zhongshan (17 samples), Zhuhai (23 samples), and Dongguan (17 samples) in September 2008 (Wei, 2010). In total 58 arthropod families were found.

Choose different distance (similarity) measures and set 1000 randomizations. The results, as indicated in Table 1, show that there is not significant difference between these cities in the arthropod composition. From p values in Table 1, the family composition of arthropods between Zhuhai and Dongguan is relatively more different.

Table 1 The *p* values for city pairs

		Zhongshan	Zhuhai	Dongguan	Guangzhou
Euclidean	Zhongshan		0.859	0.566	0.858
	Zhuhai			0.217	0.584
	Dongguan				0.887
Pearson	Zhongshan		0.855	0.554	0.864
	Zhuhai			0.207	0.565
	Dongguan				0.863
Point	Zhongshan		0.870	0.582	0.858
	Zhuhai			0.221	0.570
	Dongguan				0.846
Jaccard	Zhongshan		0.849	0.577	0.872
	Zhuhai			0.205	0.557
	Dongguan				0.856

References

Clarke KR. 1993. Non-parametric multivariate analyses of changes in community structure. Australia Journal of Ecology, 18:117-143

Damgaard C. 2011. Measuring competition in plant communities where it is difficult to distinguish individual plants. Computational Ecology and Software, 1(3): 125-137

Lüi XR. 2011. Quantitative risk analysis and prediction of potential distribution areas of common lantana (*Lantana Camara*) in China. Computational Ecology and Software, 1(1): 60-65

Rai V, Upadhyay RK, Raw SN, et al. 2011. Some aspects of animal behavior and community dynamics. Computational Ecology and Software, 1(3): 153-182

Schoenly KG, Zhang WJ. 1999. IRRI Biodiversity Software Series. V. RARE, SPPDISS, and SPPANK: programs for detecting between-sample difference in community structure. IRRI Technical Bulletin No. 5. International Rice Research Institute, Manila, Philippines

Watts MJ, Worner SP. 2011. Improving cluster-based methods for investigating potential for insect pest species establishment: region-specific risk factors. Computational Ecology and Software, 1(3): 138-145

Wei W. 2010. Biodiversity Analysis on Arthropod and Weed Communities in Paddy Rice Fields of Pearl River Delta. Master Degree Dissertation. Sun Yat-sen University, China

Zhang WJ. 2011. Constructing ecological interaction networks by correlation analysis: hints from community sampling. Network Biology, 1(2): 81-98

Zhang WJ, Chen B. 2011. Environment patterns and influential factors of biological invasions: a worldwide survey. Proceedings of the International Academy of Ecology and Environmental Sciences, 1(1): 1-14