Short Communication

A Java algorithm for non-parametric statistic comparison of network 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 29 May 2011; *Accepted 3 July 2011*; *Published online 1 September 2011*

IAEES

Abstract

A Java algorithm to statistically compare between-network structure difference was developed. In this algorithm, Euclidean distance, Manhattan distance, Chebyshov distance, and Pearson correlation were available to measure between-network difference. The algorithm was tested and applied for its effectiveness with some arthropod and weed networks.

Keywords network; structure; non-parametric statistic test; comparison; Java algorithm.

1 Introduction

The structure of network refers to the node degree, network connectance, aggregation strength (Dormann, 2011; Zhang, 2011; Zhang and Zhan, 2011), etc. We occasionally need to compare the structure difference between networks. Non-parametric statistics may be used in the difference comparison (Solow, 1993; Manly, 1997; Zhang, 2007).

In this article a Java algorithm, based on previous studies, was presented to statistically compare between-network structure difference.

2 Algorithm

The algorithm is used to compare the difference in structure composition between two networks.

Suppose that a_{ij} is the mass (or degree, etc.) of node j in network i, i=1,2,...,n; j=1,2,...,s. First, define between-network distance measures, i.e., Euclidean distances, Manhattan distances, Chebyshov distance, Pearson correlation (based distance), are as follows:

$$d_{ij}=(\sum^{s}_{k=1}(a_{ik} - a_{jk})^2/s)^{0.5}$$
$$d_{ij}=\sum^{s}_{k=1}|a_{ik} - a_{jk}|/s$$
$$d_{ij}= max_{\ k}\ |a_{ik} - a_{jk}|$$
$$d_{ij}=1-\sum^{s}_{k=1}((a_{ik} -a_{ibar})(a_{jk}-a_{jbar}))/(\sum^{s}_{k=1} (a_{ik}-a_{ibar})^2 \sum^{s}_{k=1} (a_{jk}-a_{jbar})^2)^{0.5}$$

where a_{ibar} and a_{jbar} are means of a_{ik} 's and a_{jk} 's.

If $min\ a_{ij}$<0, then let $a_{ij}=a_{ij}-min\ a_{ij}$, i=1,2,...,n; j=1,2,…,s. Suppose z_{ij} is the decimal numbers of a_{ij} if network data contain the decimal value a_{ij}, and calculate $c_{ij}=10^{Zij}$. Let $a_{ij}=a_{ij}\ max\ c_{kl}$, i=1,2,...,n; j=1,2,…,s. Through

these transformations all of the values in network data become integers which are equivalent to numbers of individuals. If no difference exists, then the distribution of individuals in networks *i* and *j* will be a result of allocating the mixed network values at random into two networks of size equal to those of the original network (Solow, 1993; Manly, 1997; Zhang, 2007). Assume that the two networks to be tested are *i* and *j*, which contain $\sum^{s}_{k=1} a_{ik}$ and $\sum^{s}_{k=1} a_{jk}$ individuals respectively. The $\sum^{s}_{k=1} a_{ik}+\sum^{s}_{k=1} a_{jk}$ individuals of the combined network are randomly reallocated into two randomized networks with $\sum^{s}_{k=1} a_{ik}$ and $\sum^{s}_{k=1} a_{jk}$ labeled individuals. Calculate the expected absolute distance between the two randomized networks and compare whether it is not less than the absolute distance between the true networks *i* and *j*. Repeat the simulation many times, calculate the number of the expected are not less than the absolute distance between *i* and *j*, and take the percentage as the *p* value. The *p* value is used to make statistical test. The threshold *p* value for test may be defined as 0.05, 0.01, etc. If the calculated *p* value is less than *p* threshold, then the structure composition of networks *i* and *j* are statistically different.

The algorithm, NetStructComp, is implemented as a Java program 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 network data file, the first row is ID numbers of nodes and the first column is ID numbers of networks.

3 Application

I chose the weed data of rice fields in four cities (networks) of Pearl River Delta, China. In total 25 plant families (nodes) were found (Wei, 2010), as indicated in Table 1.

Table 1 Abundance of plants around rice fields in four cities of China

Plant Family	Zhongshan	Zhuhai	Dongguan	Guangzhou
Gramineae	1056.9	184.6	439.3	193.6
Compositae	11.1	95	43.3	63.4
Amaranthaceae	31.1	56	93.4	49
Commelinaceae	0	52.2	14.4	1.4
Onagraceae	0	0.1	2.3	1
Urticaceae	0.3	0	11.9	3.4
Menispermaceae	0	0	0	0.1
Cyperaceae	0	0	0	26.1
Caryophyllaceae	0	0	5.3	6.7
Polygonaceae	0.4	4.2	6.9	3.8
Acanthaceae	0	0	0	0.3
Solanaceae	0.1	0	0.2	0.4
Umbelliferae	0	34.9	0	4.8
Lythraceae	0	0	0	1.6
Scrophulariaceae	0.7	3.9	1.7	2.4
Oxalidaceae	0	0	0	0.2
Chenopodiaceae	1.1	0.4	0.3	0.1
Haloragaceae	0	0	0	4.6
Campanulaceae	0	0	0	0.7
Plantaginaceae	0	0.3	0	0
Rubiaceae	0	0.1	0	0
Euphorbiaceae	0	0	0.1	0
Convolvulaceae	0	0.1	0	0
Pontederiaceae	0	0.1	0	0
Portulacaceae	24.6	0	8	0

Choose Euclidean distance measure, significance level p=0.01, and 1000 randomizations, the results are as follows:

Network pairs with significant statistic difference in structure (with p values):
(1,2)(0.0) (1,3)(0.0) (1,4)(0.0)
(2,3)(0.0) (2,4)(0.0)
(3,4)(0.0)

It is obvious that all network pairs have significant statistic difference.

Another data set is the arthropod data of nine rice fields of Pearl River Delta, China. In total 5 arthropod groups (nodes) were found (Wei, 2010), as indicated in Table 2.

Table 2 Arthropod abundance in nine rice fields

	Herbivorous Insects	Neutral Insects	Predatory Insects	Parasitic Insects	Spiders
1	42.0	0.0	5.3	3.0	8.3
2	66.4	0.0	7.9	4.3	5.7
3	298.8	0.0	10.5	3.2	10.8
4	58.1	0.0	8.9	3.1	6.9
5	50.2	0.0	6.5	3.5	4.5
6	90.6	0.0	19.6	5.6	8.0
7	53.0	0.0	10.0	3.0	7.0
8	36.1	0.1	6.9	3.2	8.2
9	40.3	0.0	8.4	2.4	11.9

Choose Euclidean distance measure, significance level p=0.01, and 1000 randomizations, the results are as follows:

(1,2)(0.0) (1,3)(0.0) (1,5)(0.0010)
(2,3)(0.0) (2,7)(0.0020) (2,8)(0.0) (2,9)(0.0)
(3,4)(0.0) (3,5)(0.0) (3,6)(0.0) (3,7)(0.0) (3,8)(0.0) (3,9)(0.0)
(4,8)(0.0) (4,9)(0.0)
(5,8)(0.0) (5,9)(0.0)
(6,8)(0.0) (6,9)(0.0)
(7,8)(0.0020) (7,9)(0.0)

The results demonstrate that, for example, rice fields 4, 5, 6 and 7 are significantly different from rice fields 8 and 9.

References

Dormann CF. 2011. How to be a specialist? Quantifying specialisation in pollination networks. Network Biology, 1(1): 1-20

Manly BFJ. 1997. Randomization, Bootstrap and Monte Carlo Methods in Biology (2nd Edition). Chapman &

Hall, London, UK

Solow AR. 1993. A simple test for change in community structure. Journal of Animal Ecology, 62: 191-193

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. 2007. Computer inference of network of ecological interactions from sampling data. Environmental Monitoring and Assessment, 124: 253-261

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

Zhang WJ, Zhan CY. 2011. An algorithm for calculation of degree distribution and detection of network type: with application in food webs. Network Biology, 1(3-4) (Accepted)