

Article

General correlation and partial correlation analysis in finding interactions: with Spearman rank correlation and proportion correlation as correlation measures

WenJun Zhang¹, Xin Li²

¹School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China; International Academy of Ecology and Environmental Sciences, Hong Kong

²College of Plant Protection, Northwest A & F University, Yangling 712100, China; Yangling Institute of Modern Agricultural Standardization, Yangling 712100, China

E-mail: zhwj@mail.sysu.edu.cn, wzhang@iaees.org, lixin57@hotmail.com

Received 19 November 2015; Accepted 25 December 2015; Published online 1 December 2015



Abstract

Between-taxon interactions can be detected by calculating the sampling data of taxon \times sample type. In present study, Spearman rank correlation and proportion correlation are chosen as the general correlation measures, and their partial correlations are calculated and compared. The results show that for Spearman rank correlation measure, in all predicted candidate direct interactions by partial correlation, about 16.77% (x , 0~45.4%) of them are not successfully detected by Spearman rank correlation. In all predicted interactions by Spearman rank correlation, 47.56% (y , 0~100%) of them are undeterministic interactions, i.e., not successfully detected by partial correlation. In all predicted interactions by Spearman rank correlation, 53.45% (z , 0~100%) of them are candidate interactions, i.e., successfully detected by partial correlation. The regression relationship between Spearman rank correlation (r) and its partial correlation (pr) is $pr=0.0102+0.1085r$ ($R^2=0.0181$, $p<0.00001$, $n=1004$). For proportion correlation measure, in all predicted candidate interactions by partial correlation, about 6.82% (x , 0~28.64%) of them are not successfully detected by proportion correlation. In all predicted interactions by proportion correlation, 72.24% (y , 28.01~100%) of them are undeterministic interactions. In all predicted interactions by proportion correlation, 27.76% (z , 0~71.99%) of them are candidate interactions. The regression relationship between proportion correlation and its partial correlation is $pr=0.07+0.0592r$ ($R^2=0.0213$, $p<0.00001$, $n=1447$). The proportion of missed (x), mis-predicted (y) and precisely predicted candidate direct interactions (z) by general correlation analysis increases, increases, and decreases with the number of taxa respectively. Relationships between general correlation (r) and partial correlation (pr) mean that indirect interactions increase mean interaction strength of taxa. The precisely predicted (z) candidate direct interactions by Spearman rank correlation and proportion correlation analysis are not necessarily those with the highest Spearman rank correlations and proportion correlations. Jointly using correlation and partial correlation measures to analyze various interactions is the most reliable choice. Candidate direct interactions detected by both correlation and partial correlation measures should be the most focused interactions, seconded by those interactions detected by partial correlation only and by correlation only.

Keywords false and true direct interactions; Spearman rank correlation; proportion correlation; partial correlation; statistic test.

Network Biology
ISSN 2220-8879
URL: <http://www.iaees.org/publications/journals/nb/online-version.asp>
RSS: <http://www.iaees.org/publications/journals/nb/rss.xml>
E-mail: networkbiology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

In earlier studies (Zhang, 2007, 2011, 2012a, 2012b, 2015; Zhang and Li, 2015), methodology for detecting interactions by correlation analysis of sampling data were proposed. In present study, Spearman rank correlation and proportion correlation are chosen as the general correlation measures, and their partial correlations are calculated and compared, in order to provide basis for predicting interactions.

2 Material and Methods

Spearman rank correlation and proportion correlation are measures to reflect the rank and proportion dependence between two taxa (Zhang, 2015). A statistically significant Spearman rank correlation (proportion correlation) represents a direct or indirect interaction between two taxa. Partial correlation is based on Spearman rank correlation (proportion correlation). It helps to eliminate the indirect effects produced by the remaining taxa. A statistically significant partial correlation represents a candidate direct interaction between two taxa (Zhang, 2007, 2011, 2012a, 2012b, 2015; Zhang and Li, 2015). In present study, we treated the linear interactions, predicted by partial correlation, as candidate direct interactions. Methods followed Zhang (2015), and the data on biological networks were from Zhang (2011; Fig. 1). These biological networks are different in countries, years, seasons, types of taxa, and number of taxa. Therefore we can expect the wide representativeness of conclusions drawn from them.



Fig. 1 Some of the rice arthropods. Pictures were taken by the author on July 29, 2015, from Kaili, Guizhou, China.

3 Results

3.1 Error estimation

All results are listed in Tables 1 and 2. For Spearman rank correlation measure, in all predicted candidate direct interactions by partial correlation, about 16.77% (x , 0~45.4%) of them are not successfully detected by Spearman rank correlation. In all predicted interactions by Spearman rank correlation, 47.56% (y , 0~100%) of them are undeterministic interactions, i.e., not successfully detected by partial correlation. In all predicted

interactions by Spearman rank correlation, 53.45% (z , 0~100%) of them are candidate interactions, i.e., successfully detected by partial correlation (Table 1). The regression relationship between Spearman rank correlation (r) and its partial correlation (pr) is

$$pr=0.0102+0.1085r, R^2=0.0181, p<0.00001, n=1004$$

The regression relationships between number of taxa (N) and proportion of missed (x), proportion of mis-predicted (y) and proportion of precisely predicted (z) are as follows

$$x=7.1227+0.7875N, R^2=0.2373, p=0.1082, n=12$$

$$y=-8.9854+4.6155N, R^2=0.808, p=0.0001, n=12$$

$$z=108.9854-4.6155N, R^2=0.808, p=0.0001, n=12$$

For proportion correlation measure, in all predicted candidate direct interactions by partial correlation, about 6.82% (x , 0~28.64%) of them are not successfully detected by proportion correlation (Table 2). In all predicted interactions by proportion correlation, 72.24% (y , 28.01~100%) of them are undeterministic interactions, i.e., not successfully detected by partial correlation. In all predicted interactions by proportion correlation, 27.76% (z , 0~71.99%) of them are candidate direct interactions, i.e., successfully detected by partial correlation (Table 1). The regression relationship between proportion correlation (r) and its partial correlation (pr) is

$$pr=0.07+0.0592r, R^2=0.0213, p<0.00001, n=1447$$

Table 1 Comparison of results of Spearman rank correlation (SRC) and partial Spearman correlation (Partial SRC)

Network ID (Data set)	No. Taxa (N)	No. SS Yes Partial SRC but SS Not SRC (SSN)	$x=SSN/(SN+SY)$ (%)	No. SS Yes SRC but SS Not Partial SRC (SPN)	$y=SPN/(SPN+SY)$ (%)	No. SS Yes SRC & SS Yes Partial SRC (SY)	$z=SY/(SPN+SY)$ (%)
CN-06sep	4	0	0	0	0	2	100
CN-06sep	4	0	0	1	33.33	2	66.67
CN-06Oct	4	0	0	0	0	2	100
CN-06Oct	4	0	0	0	0	1	100
PH-Mar	21	3	30	81	92.05	7	7.95
PH-Apr	20	4	36.36	94	93.07	7	6.93
PH-Sep	21	2	25	110	94.83	6	5.17
PH-Oct	21	2	22.22	112	94.12	7	5.88
PH-Apr	7	1	33.33	4	66.67	2	33.33
PH-Sep	7	1	16.67	0	0	5	100
PH-Oct	7	2	28.57	0	0	5	100
CN-06Oct	27	1	9.09	283	96.59	10	3.41
Mean			16.77		47.56		53.45
\pm ($p\leq 0.05$)			28.63		90.92		90.92

No. SS Yes Partial SRC but SS Not SRC: Total No. of statistically significant Partial SRC ($p\leq 0.01$) but statistically not significant SRC ($p\leq 0.01$); No. SS Yes SRC but SS Not Partial SRC: Total No. of statistically significant SRC ($p\leq 0.01$) but statistically not significant Partial SRC ($p\leq 0.01$); No. SS Yes SRC & SS Yes Partial SRC: Total No. of both statistically significant SRC ($p\leq 0.01$) and statistically significant Partial SRC ($p\leq 0.01$).

Table 2 Comparison of results of proportion correlation (PC) and partial proportion correlation (Partial PC).

Network ID (Data set)	No. Taxa (<i>N</i>)	No. SS Yes Partial PC but SS Not PC (SSN)	$x=SSNA/($ $SSN+SY)$) (%)	No. SS Yes PC but SS Not Partial PC (SPN)	$y=SPN/(SP$ $N+SY)$ (%)	No. SS Yes PC & SS Yes Partial PC (SYY)	$z=SY/(SP$ $N+SY)$ (%)
CN-06sep	4	0	0	1	33.33	2	66.67
CN-06sep	4	0	0	2	50	2	50
CN-06Oct	4	0	0	2	50	2	50
CN-06Oct	4	0	0	2	50	2	50
PH-Mar	21	0	0	78	92.86	6	7.14
PH-Apr	20	1	12.5	63	90	7	10
PH-Sep	21	0	0	67	91.78	6	8.22
PH-Oct	21	0	0	61	95.31	3	4.69
CN-06Sep	23	2	16.67	32	76.19	10	23.81
CN-06Oct	23	1	12.5	39	84.78	7	15.22
CN-06Oct	27	5	33.33	41	80.39	10	19.61
Mean			6.82		72.24		27.76
\pm ($p \leq 0.05$)			21.82		44.23		44.23

No. SS Yes Partial PC but SS Not PC: Total No. of statistically significant Partial PC ($p \leq 0.01$) but statistically not significant PC ($p \leq 0.01$); No. SS Yes PC but SS Not Partial PC: Total No. of statistically significant PC ($p \leq 0.01$) but statistically not significant Partial PC ($p \leq 0.01$); No. SS Yes PC & SS Yes Partial PC: Total No. of both statistically significant PC ($p \leq 0.01$) and statistically significant Partial PC ($p \leq 0.01$).

The regression relationships between number of taxa (N) and proportion of missed (x), proportion of mis-predicted (y) and proportion of precisely predicted (z) are as follows

$$x = -4.6383 + 0.7327N, R^2 = 0.3991, p = 0.0371, n = 11$$

$$y = 39.3633 + 2.1026N, R^2 = 0.7994, p = 0.0002, n = 11$$

$$z = 60.6367 - 2.1026N, R^2 = 0.7994, p = 0.0002, n = 11$$

The relationships between general correlation (r) and partial correlation (pr) mean that indirect interactions in a network increase mean interaction strength of taxa.

3.2 Precisely predicted (z) candidate direct interactions by partial correlation analysis

Our results showed that the precisely predicted (z) candidate direct interactions by Spearman rank correlation and proportion correlation analysis are not necessarily those with the highest Spearman rank correlations and proportion correlations. For example, the predicted interactions by Spearman rank correlation analysis (and Spearman rank correlations) of network PH-Mar (taxon: functional group; in total 21 functional groups) are indicated in Table 3, of which the interactions in bold are candidate direct interactions.

Therefore we should not try to choose a portion (e.g., z , as calculated above) of predicted interactions with the greatest Spearman rank correlations or proportion correlations as candidate direct interactions.

4 Discussion

The proportions of missed (x), mis-predicted (y), and precisely predicted (z) candidate direct interactions by Pearson linear correlation and proportion correlation analysis have significant biological meaning, as discussed in Zhang and Li (2015).

Table 3 Statistically significant Spearman rank correlations and true interactions determined by partial correlations (in bold).

1	3	0.612	11	13	0.4819	15	17	-0.385	11	20	0.4651
2	3	0.4768	1	14	0.5752	6	18	0.3596	14	20	0.4801
3	4	0.539	2	14	0.5148	8	18	0.375	19	20	0.4801
3	5	0.353	3	14	0.8515	11	18	0.4396	1	21	0.5759
3	6	0.4888	4	14	0.6124	12	18	-0.354	2	21	0.4593
4	6	0.3923	5	14	0.4823	14	18	0.538	3	21	0.8515
2	7	0.4828	6	14	0.5087	1	19	0.531	4	21	0.6599
3	7	0.4247	7	14	0.5126	2	19	0.4497	5	21	0.5187
3	8	0.4196	8	14	0.5252	3	19	0.8515	6	21	0.5245
7	8	0.4959	9	14	0.5537	4	19	0.6591	7	21	0.4573
3	9	0.429	10	14	0.5179	5	19	0.564	8	21	0.4186
3	10	0.4391	11	14	0.9	6	19	0.5304	9	21	0.4561
1	11	0.531	12	14	0.3675	7	19	0.4268	10	21	0.4705
2	11	0.4922	13	14	0.5344	8	19	0.4543	11	21	0.9
3	11	0.8515	3	15	0.4342	9	19	0.4561	12	21	0.4468
4	11	0.6591	11	15	0.521	10	19	0.4979	13	21	0.5347
5	11	0.5732	14	15	0.456	11	19	0.9	14	21	0.9
6	11	0.5087	3	16	0.5434	12	19	0.4717	15	21	0.4066
7	11	0.4835	4	16	0.4235	13	19	0.4819	16	21	0.5875
8	11	0.416	11	16	0.5455	14	19	0.9	17	21	0.4809
9	11	0.507	13	16	0.4006	15	19	0.5251	18	21	0.4356
10	11	0.4704	14	16	0.483	16	19	0.483	19	21	0.9
3	12	0.5468	11	17	0.459	17	19	0.3706	20	21	0.5105
11	12	0.4567	12	17	-0.491	18	19	0.3696			
3	13	0.454	14	17	0.5065	3	20	0.3954			

As done in Zhang and Li (2015), we suggest jointly using general correlation and partial correlation to analyze various interactions. Candidate direct interactions detected by both general correlation measures should be the most focused interactions, seconded by those interactions detected by partial correlation only and by general correlation only.

Acknowledgment

We are thankful to the support of Discovery and Crucial Node Analysis of Important Biological and Social Networks (2015.6-2020.6), from Yangling Institute of Modern Agricultural Standardization, High-Quality Textbook *Network Biology* Project for Engineering of Teaching Quality and Teaching Reform of Undergraduate Universities of Guangdong Province (2015.6-2018.6), from Department of Education of Guangdong Province, and Project on Undergraduate Teaching Reform (2015.7-2017.7), from Sun Yat-sen University, China.

References

- 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. 2012a. *Computational Ecology: Graphs, Networks and Agent-based Modeling*. World Scientific, Singapore
- Zhang WJ. 2012b. How to construct the statistic network? An association network of herbaceous plants constructed from field sampling. *Network Biology*, 2(2): 57-67
- Zhang WJ, Li X. 2015. Linear correlation analysis in finding interactions: Half of predicted interactions are undeterministic and one-third of candidate direct interactions are missed. *Selforganizology*, 2(3): 39-45
- Zhang WJ. 2015. Calculation and statistic test of partial correlation of general correlation measures. *Selforganizology*, 2(4): 65-77