Article

Linear correlation analysis in finding interactions: Half of predicted interactions are undeterministic and one-third of candidate direct interactions are missed

WenJun Zhang[1], **Xin Li**[2]

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

[2]College 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, lixin57@hotmail.com

Received 16 October 2014; *Accepted 25 June 2015*; *Published online 1 September 2015*

Abstract

An ecological network can be constructed by calculating the sampling data of taxon × sample type. A statistically significant Pearson linear correlation means an indirect or direct linear interaction between two taxa, and a statistically significant partial correlation based on Pearson linear correlation, due to elimination of indirect effects of other taxa, means a candidate direct interaction between two taxa. People always use Pearson linear correlation to find interactions. However, some undeterministic interactions may be found and some candidate direct interactions may be missed when using this method. The results show that partial linear correlation (y) is approximately half of the Pearson linear correlation (x) (y=-0.0064+0.4785x, r^2=0.173, p<0.00001, n=1447), which means that indirect interactions increase mean interaction strength of taxa in the network. In all predicted interactions by partial linear correlation, about 34.35% (x, 0~100%) (i.e., one-third) of them are not successfully detected by linear correlation. In all predicted interactions by Pearson linear correlation, 50.58% (y, 0~100%) (i.e., half) of them are undeterministic interactions, i.e., not successfully detected by partial linear correlation, and 49.42% (z, 0~100%) (i.e., half) of them are candidate direct interactions, i.e., successfully detected by partial linear correlation also. The proportion of missed (x), mis-predicted (y) and precisely predicted candidate direct interactions (z) by Pearson linear correlation analysis decreases (r=-0.49, p=0.07), increases (r=0.48, p=0.08), and decreases (r=-0.48, p=0.08) slightly with the number of taxa (m) respectively. Results show that the precisely predicted (z) candidate direct interactions by Pearson linear correlation analysis are not necessarily those with the highest Pearson linear correlations. We should not try to choose a portion (e.g., 49.42% (z)) of predicted interactions with the greatest Pearson linear correlations as candidate direct interactions. We suggest jointly using Pearson linear correlation and partial linear correlation to analyze various interactions. Candidate direct interactions detected by both linear correlation measures should be the most focused interactions, seconded by those interactions detected by partial linear correlation only and by Pearson linear correlation only.

Keywords direct interactions; indirect interactions; Pearson linear correlation; partial linear correlation.

Selforganizology
ISSN 2410-0080
URL: http://www.iaees.org/publications/journals/selforganizology/online-version.asp
RSS: http://www.iaees.org/publications/journals/selforganizology/rss.xml
E-mail: selforganizology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

In a series of earlier studies (Zhang, 2007, 2011, 2012a, 2012b), methodology for constructing ecological networks by correlation analysis of community sampling data have been proposed. We mentioned that a statistically significant Pearson linear correlation means an indirect or direct interaction between two taxa, and a statistically significant partial (net, or pure) correlation based on Pearson linear correlation means a candidate direct interaction between two taxa. For the studies of ecological communities and ecosystems, interactions refer to predation, parasitism, competition, amensalism, mutualism, protocooperation, commensalism, etc. Two taxa may interact by acting to the same resource, or by changing the environment of opposite sides, etc. An interaction means a dependency relationship in state changes of two taxa (direct interaction). Conversely, a seeming dependency relationship in state changes of two taxa does not necessarily mean an interaction (indirect interaction).

People always use Pearson linear correlation to find interactions (Goh, et al., 2000; Pazos and Valencia, 2001; Tu, 2006). However, some undeterministic interactions may be found and some candidate direct interactions may be missed when using this method, as pointed out by Zhang (2011). In present study, we tried to find the error of predicting interactions with correlation analysis.

2 Material and Methods

A network is globally the linear network, quasi-linear network or nonlinear network. Furthermore, a network changed in a local domain (a short time, a small extent) can be approximated as a linear network (Zhang, 2011, 2012a, 2012b), i.e., in the local domain, all between-node (or -taxon, -component, etc) changes are treated as linear ones, i.e., suppose x_i is the state of node i, i=1, 2, …, m, then we have

$$dx_i(t)/dt=a_{ij}\, dx_j(t)/dt \qquad i, j=1, 2, \ldots, m$$

or

$$dx_i(l)/dl=a_{ij}\, dx_j(l)/dl \qquad i, j=1, 2, \ldots, m$$

where t: time; l: space length; a_{ij}: constants, i, j=1, 2, …,m, and a_{ii}=1, i=1, 2, …, m. In these situations, linear correlation measures can thus be used. Pearson linear correlation is a measure to reflect the linear dependence between two taxa. A statistically significant Pearson linear correlation represents a direct or indirect linear interaction between two taxa. Partial (net, or pure) linear correlation is based on Pearson linear correlation. It has eliminated the indirect effects produced by the remaining taxa. A statistically significant partial linear correlation represents a candidate direct linear interaction between two taxa (Zhang, 2007, 2011, 2012a, 2012b). In present study, we treated the linear interactions, predicted by partial linear correlation, as candidate direct interactions.

The following are Matlab codes for calculation and statistic test of Pearson linear correlation and partial linear correlation, and for finding interactions:

```
%Reference: 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
% X is m*n raw data matrix. m: number of taxa; n: number of samples.
str=input('Input the file name of raw data matrix (e.g., raw.txt, raw.xls, etc. The file has m rows (taxa) and n columns
(samples)): ','s');
X=load(str);
sig=input('Input significance level(e.g., 0.01): ');
```

```
dim=size(X);
m=dim(1); n=dim(2);
r=corr(X');
disp('Correlation matrix')
r
tvalues=abs(r)./sqrt((1-r.^2)/(n-2));
alpha=(1-tcdf(tvalues,n-2))*2;
sigmat=alpha<sig;
sigmat=sigmat.*r-eye(m);
sigmatr=sigmat;
disp('Pairs with statistically significant correlation')
if (sigmat~=ones(m))
[pairx,pairy,rvalues]=find(sigmat);
temp1=pairx; temp2=pairy;
pairxs=pairx(temp1<temp2);
pairys=pairy(temp1<temp2);
rvaluess=rvalues(temp1<temp2);
PairsAndCorrelations=[pairxs pairys rvaluess]
else
disp('No significant pairs')
end
inversr=inv(r);
for i=1:m-1; for j=i+1:m; parr(i,j)=-inversr(i,j)/sqrt(inversr(i,i)*inversr(j,j));end;end;
for i=1:m-1; for j=i+1:m; parr(j,i)=parr(i,j);end;end;
for i=1:m; parr(i,i)=1;end;
disp('Partial correlation matrix')
parr
if (n>m)
tvalues=abs(parr)./sqrt((1-parr.^2)/(n-m));
alpha=(1-tcdf(tvalues,n-m))*2;
else
disp('The number of samples is not enough to support the required statistic test (DF=n-m) of partial correlations. Here use the statistic test with DF=n-2 (not recommended). Please input the proportion of statistically significant pairs based on DF=n-m vs. statistically significant pairs based on DF=n-2 (y, %) as the following. The estimation formula, y=88.748exp(-0.045m), is suggested for use (Zhang WJ. 2015. Selforganizology, 2(4): 55-67). If it is hard to be estimated, the full percent, 100, can be input. ')
y=input('Input the proportion (a value between 0 and 100): ')
tvalues=abs(parr)./sqrt((1-parr.^2)/(n-2));
alpha=(1-tcdf(tvalues,n-2))*2;
end
sigmat=alpha<sig;
sigmat=sigmat.*parr-eye(m);
if (n<=m) threshr=rrank(sigmat,y); sigmat=sigmat>=threshr; sigmat=sigmat.*parr; end;
sigmatparr=sigmat;
disp('Pairs with statistically significant partial correlation')
if (sigmat~=ones(m))
[pairx,pairy,rvalues]=find(sigmat);
temp1=pairx; temp2=pairy;
pairxs=pairx(temp1<temp2);
pairys=pairy(temp1<temp2);
```

```
rvaluess=rvalues(temp1<temp2);
PairsAndPartialCorrelations=[pairxs pairys rvaluess]
else
disp('No significant pairs')
end
x=sigmatparr & (~sigmatr);
y=(~sigmatparr) & sigmatr;
z=sigmatparr & sigmatr;
for i=1:3;
switch i
    case 1
        mat=x; s='Significant partial correlation but insignificant linear correlation';
    case 2
        mat=y; s='Significant linear correlation but insignificant partial correlation';
    case 3
        mat=z; s='Significant both partial correlation and correlation';
end;
[pairx,pairy]=find(mat);
temp1=pairx; temp2=pairy;
pairxs=pairx(temp1<temp2);
pairys=pairy(temp1<temp2);
disp([s])
SignificantPairs=[pairxs pairys]
end;
```

The M function file, rrank.m, is as the following:

```
function threshr = rrank(mat,percent)
dim=size(mat); m=dim(1);
len=(m*m-m)/2;
vec=zeros(1,len);
n=0;
for i=1:m-1;    for j=i+1:m;
if (mat(i,j)~=0) n=n+1; vec(n)=mat(i,j); end;
end; end;
num=round(percent/100*n);
vecc=sort(vec,'descend');
if (num~=0) threshr=vecc(num); else threshr=1;
end;
```

Data of various biological networks were obtained from that of Zhang (2011). These biological networks are different in countries, years, seasons, types of taxa, and number of taxa. Therefore we expect the wide representativeness of conclusions drawn from them.

3 Results

3.1 Relationship between partial linear correlation and Pearson linear correlation

Partial linear correlation (y) is approximately half of the Pearson linear correlation (x), as indicated by

$$y=-0.0064+0.4785x,\ r^2=0.173,\ p<0.00001,\ n=1447$$

This relationship means that indirect interactions increase mean interaction strength of taxa in the network.

3.2 Error estimation

Our results are listed in Table 1. In all predicted candidate interactions by partial linear correlation, about 34.35% (*x*, 0~100%) of them are not successfully detected by linear correlation. In all predicted interactions by Pearson linear correlation, 50.58% (*y*, 0~100%) of them are undeterministic interactions, i.e., not successfully detected by partial linear correlation. In all predicted interactions by Pearson linear correlation, 49.42% (*z*, 0~100%) of them are candidate direct interactions, i.e., successfully detected by partial linear correlation also (Fig. 1).

It is found that the Pearson linear correlations between *N* and *x*, *y*, *z* are -0.49 (*p*=0.07), 0.48 (*p*=0.18) and -0.48 (*p*=0.18), respectively. The proportion of missed (*x*), mis-predicted (*y*) and precisely predicted (*z*) candidate direct interactions by Pearson linear correlation analysis decreases, increases, and decreases with the number of taxa respectively. However, statistically there is not significant linear dependency between number of taxa (*m*) and these percentage indices.

Table 1 Comparison of results of Pearson linear correlation (PLC) and partial linear correlation (Partial PLC)

Network ID (Data set)	No. Taxa (*m*)	. No. Partial PLC> PLC	No. Absolute Partial PLC> Absolute PLC(*NAP*)	*NAP*/ *N* (%)	No. SS Yes Partial PLC but SS Not PLC (*SSN*)	***x*=*SSN*/(*SSN*+*SYY*) (%)**	No. SS Yes PLC but SS Not Partial PLC (*SPN*)	***y*=*SPN*/(*SPN*+*SYY*) (%)**	No. SS Yes PLC & SS Yes Partial PLC (*SYY*)	***z*=*SYY*/(*SPN*+*SYY*) (%)**
CN-06sep	4	0	0	0.00	1	**100**	1	**50.00**	1	**50.00**
CN-06sep	4	2	3	50.00	2	**100**	0	**0.00**	2	**100.00**
CN-06Oct	4	2	3	50.00	2	**100**	1	**50.00**	1	**50.00**
PH-Mar	21	99	126	60.00	0	**0**	6	**66.67**	3	**33.33**
PH-Apr	20	57	106	55.79	4	**33.33**	14	**63.64**	8	**36.36**
PH-Sep	21	70	122	58.10	1	**25**	10	**76.92**	3	**23.08**
PH-Oct	21	98	100	47.62	0	**0**	8	**72.73**	3	**27.27**
PH-Mar	7	10	12	57.14	1	**50**	0	**0.00**	1	**100.00**
PH-Apr	7	4	6	28.57	0	**0**	5	**71.43**	2	**28.57**
PH-Sep	7	5	7	33.33	0	**0**	5	**83.33**	1	**16.67**
PH-Oct	7	11	13	61.90	0	**0**	0	**0.00**	2	**100.00**
CN-06Sep	23	120	174	68.77	3	**25**	10	**52.63**	9	**47.37**
CN-06Oct	23	116	127	50.20	1	**14.29**	10	**62.50**	6	**37.50**
CN-06Oct	27	168	248	70.66	5	**33.33**	14	**58.33**	10	**41.67**
Mean						**34.35**		**50.58**		**49.42**
± (*p*≤0.05)						**77.76**		**58.19**		**58.19**

No. SS Yes Partial PLC but SS Not PLC: Total No. of statistically significant Partial PLC (*p*≤0.01) but statistically not significant PLC (*p*≤0.01); No. SS Yes PLC but SS Not Partial PLC: Total No. of statistically significant PLC (*p*≤0.01) but statistically not significant Partial PLC (*p*≤0.01); No. SS Yes PLC & SS Yes Partial PLC: Total No. of both statistically significant PLC (*p*≤0.01) and statistically significant Partial PLC (*p*≤0.01).

3.3 Precisely predicted (*z*) candidate direct interactions by Pearson linear correlation analysis

Our results showed that the precisely predicted (*z*) candidate direct interactions by Pearson linear correlation analysis are not necessarily those with the highest Pearson linear correlations. For example, the predicted interactions by Pearson linear correlation analysis (and Pearson linear correlations in parentheses) of network CN-06Sep (taxon: family) are as follows, of which italic interactions are candidate direct interactions:

(1,2)(0.3962)
(2,3)(0.4021)
(3,4)(0.3877)
(4,10)(0.3531) *(4,16)(0.4379)* *(4,18)(0.4954)*

(5,12)(0.4095) *(5,13)(0.6502)*
(6,17)(0.43)
(8,11)(0.4017) (8,12)(0.5605)
(10,12)(0.3611) (10,18)(0.4883)
(11,12)(0.3784) (11,18)(0.451)
(12,13)(0.4703)
(14,22)(0.4201)
(16,21)(0.3541)
(17,20)(0.43)

Therefore we should not try to choose a portion (e.g., 49.42% (*z*), as calculated above) of predicted interactions with the greatest Pearson linear correlations as candidate direct interactions.

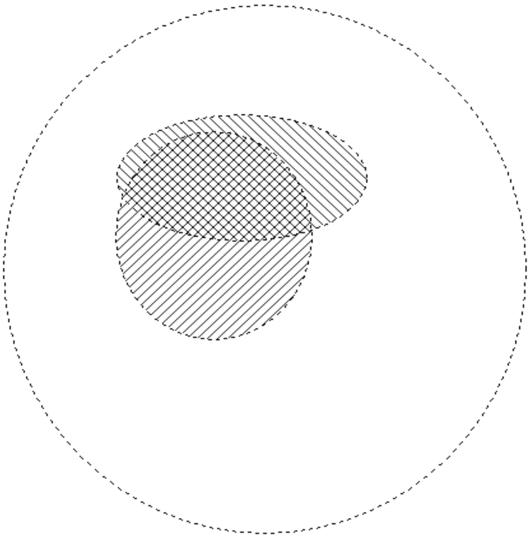

Fig. 1 Illustration of proportion of maximal possible direct interactions (inside outer circle), direct interactions detected by both linear correlation measures (▩, 49.42%), direct interactions detected by partial linear correlation but not Pearson linear correlation (▧, 34.35%), and direct interactions detected by Pearson linear correlation but not partial linear correlation (▨, 50.58%).

4 Discussion

The proportions of missed (*x*), mis-predicted (*y*), and precisely predicted (*z*) candidate direct interactions by Pearson linear correlation analysis have significant biological meaning, for example, if we want to detect candidate direct interactions between species by Pearson linear correlation analysis, a false conclusion may sometimes be drawn (*y*). For example, Tu (2006) found that some true interactions (i.e., the interactions confirmed by experiments; A candidate direct interaction is a true interaction if it is confirmed by experiments) between proteins have not statistically significant Pearson linear correlation. In medical science (biological control, biodiversity conservation), assume there is a true direct interaction between two proteins (a predator and a prey, two species), *A* and *B*. We want to develop a medicine (release *A*, release *A*) to affect *A* or *B* (to control *B*, to balance *B*) directly, and hope that the measure will make *A* affect (control, balance) *B*. However, we know that *A* may likely not affect (control, balance) *B* due to the proportion *x* (medicine failure, failure of biological control, failure of balancing plan).

According to Tu (2006), the proportions of missed (*x*) and precisely predicted (*z*) true direct interactions by Pearson linear correlation analysis in small-cell lung cancer (SCLC; *m*=19 protein sequences) were 25%

and 75%, respectively. For non-small-cell lung cancer (NSCLC; m=92 protein sequences), the proportions of missed (x) and precisely predicted (z) true direct interactions by Pearson linear correlation analysis were 42.5% and 57.5%, respectively. The results (means of SCLC and NSCLC: x=33.8%, z=66.3%) are of better coincident with our conclusions (x≈34.35%; y≈49.42%). The results of Tu indicated that true direct interactions by Pearson linear correlation analysis are not necessarily those with the highest Pearson linear correlations, which confirmed our conclusion also.

In ecological networks (systems), the phenomena above are popular. For example, it is well known that ladybird *Coccinella septempunctata* is the natural enemy of cotton aphid *Aphis gossypii* Glover, and thus there is a true (direct) interaction between the two species. However, in some cotton fields, the lady bird lacks of prey species and it mainly feeds on the cotton aphid, and the Pearson linear correlation between them is thus likely significant (the interaction belongs to z). In other cotton fields, the ladybird may own many available prey species and the Pearson linear correlation between them is thus likely insignificant (the interaction belongs to x). In this situation, the ladybird, as a natural enemy of cotton aphid, is ineffective.

In present study, we used the significance level p<0.01. To avoid missing candidate interactions as possible as, the significance level can be adjusted to a reasonable value, for example, p<0.05.

In general, we suggest jointly using Pearson linear correlation and partial linear correlation to analyze various interactions. Candidate direct interactions detected by both linear correlation measures should be the most focused interactions, which have the most significant biological meaning, seconded by those interactions detected by partial linear correlation only and by Pearson linear correlation only.

Acknowledgment

We are thankful to the support of Discovery and Crucial Node Analysis of Important Biological Networks (2015.6-2020.6), from Yangling Institute of Modern Agricultural Standardization, and 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, China.

References

Goh CS, Bogan AA, Joachimiak M, et al. 2000. Co-evolution of proteins with their interactions partners. Journal of Molecular Biology, 299: 283-293

Pazos F, Valencia A. 2001. Similarity of phylogenetic trees as indicators of protein-protein interaction. Protein Engineering, 14: 609-614

Tu WJ. 2006. Protein-protein interactions of lung cancer related proteins. MSc Thesis, Sun Yat-sen University, Guangzhou, 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. 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): 57-68

Zhang, WJ. 2015. Calculation and statistic test of partial correlation of general correlation measures. Selforganizology, 2(4): 55-67