Article

A hierarchical method for finding interactions: Jointly using linear correlation and rank correlation analysis

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 20 October 2015; *Accepted 5 November 2015*; *Published online 1 December 2015*

Abstract
In the earlier studies, I pointed out that a network changed in a local domain can be approximated as a linear network, i.e., all between-node (or -taxon, -component, etc) changes in the local domain are treated as linear ones and Pearson linear correlation measure can be used. For a little wider domain, the quasi-linear measure, Spearman rank correlation can be used also. In present study, I jointly use Pearson linear correlation measure and Spearman rank correlation measure and their partial correlations to find interactions. First, I define some hierarchical principles for finding interactions. Reliability levels are then defined using set operations. The full algorithm and Matlab codes for finding interactions are given.

Keywords partial correlation; correlation measure; Pearson linear correlation; Spearman rank correlation; algorithm; set operation; statistic test; interaction finding.

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 our earlier studies (Zhang, 2007, 2011, 2012a, 2012b, 2015, 2016; Zhang and Li, 2015a, b), we have proposed a series of methods for finding interactions by correlation analysis of sampling data. For the studies of ecological communities and ecosystems, there are generally two types of interactions, direct interactions and indirect interactions. Direct 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). Definition of interactions in other biological networks (metabolic regulatory networks, cancer networks, etc) can be found in related references.

In present study, I jointly use Pearson linear correlation measure and Spearman rank correlation measure and their partial correlations for finding interactions. Some hierarchical principles for finding interactions are defined. Reliability levels are defined using set operations. The full algorithm and Matlab codes for finding interactions are given.

2 Method

In the earlier studies (Zhang and Li, 2015a, b; Zhang, 2015), we pointed out that a network changed in a local domain (a short time, a small extent) can be approximated as a linear network, i.e., in the local domain, all between-node (or -taxon, -component, etc) changes are treated as linear ones, and Pearson linear correlation measure can be used. For a little wider domain, the quasi-linear measure, Spearman rank correlation (Spearman, 1904; Schoenly and Zhang, 1999; Zhang, 2011, 2012a, 2012b, 2015, 2016) can be used also. Therefore, we can jointly use Pearson linear correlation measure and Spearman rank correlation measure, and their partial correlations to find interactions.

Here I define the sets of interactions

A: Interactions detected by statistically significant partial correlations for Pearson linear correlation measure;

B: Interactions detected by statistically significant partial correlations for Spearman rank correlation measure;

C: Interactions detected by statistically significant Pearson linear correlations;

D: Interactions detected by statistically significant Spearman rank correlations.

When finding interactions, some principles are followed: (1) Pearson linear correlation measure is more reliable than Spearman rank correlation measure; (2) partial correlation measure is more reliable than correlation measure; (3) jointly use of Pearson linear correlation measure and Spearman rank correlation measure is more reliable than the single use of two measures, and (4) jointly use partial correlation measure and correlation measure is more reliable than the single use.

According to the above principles and the rules of set operation, I defined reliability levels (basically, from the set of most reliable interactions, L1, to the sets of most unreliable interactions, L61 and L62) as follows

L1: A∩B∩C∩D;	(candidate direct interactions)
L21: A∩C; L22: B∩D;	(candidate direct interactions)
L3: A∩B;	(candidate direct interactions)
L41: A; L42: B;	(candidate direct interactions)
L5: C∩D;	(indirect/direct interactions)
L61: C; L62: D.	(indirect/direct interactions)

Known the raw data is a matrix, *X*, with *m* rows, i.e., *m* attributes (taxa, proteins, genes, etc), and *n* columns, i.e., *n* samples. The Matlab codes of the full algorithm above are listed as follows:

```
% Reference: Zhang WJ. 2015. A hierarchical method for finding interactions: Jointly using linear correlation and
% rank correlation analysis. Network Biology, 5(4): 137-145
% X is the m*n raw data matrix. m: number of attributes (taxa, proteins, genes, etc); 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);
if (n<=m)
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. For Pearson linear correlation measure, the estimation formula, y1=88.748 exp(-0.045m), is suggested for use, and for Spearman rank correlation measure, y2=120.687exp(-0.045m), is suggested, where m is the number of attributes (taxa, proteins, genes, etc). If it is hard to be estimated, the full percent, 100, can be input. ')

```
y1=input('Input the proportion for Pearson linear correlation measure (a value between 0 and 100): ')
y2=input('Input the proportion for Spearman rank correlation measure (a value between 0 and 100): ')
end;
r=corr(X');
disp('Pearson linear correlation matrix')
r
for i=1:m-1; for j=i+1:m; spr(i,j)=spearman(X(i,:)',X(j,:)');end;end;
for i=1:m-1; for j=i+1:m; spr(j,i)=spr(i,j);end;end;
for i=1:m; spr(i,i)=1;end;
disp('Spearman rank correlation matrix')
spr
for k=1:2;
if (k==1) rr=r; else rr=spr; end;
inverse=inv(rr);
for i=1:m-1; for j=i+1:m; parr(i,j)=-inverse(i,j)/sqrt(inverse(i,i)*inverse(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;
if (k==1)
disp('Partial correlation matrix for Pearson linear correlation')
peparr=parr;
peparr
else
disp('Partial correlation matrix for Spearman rank correlation')
spparr=parr;
spparr
end;
end;
for k=1:2;
if (k==1)
tvalues=abs(r)./sqrt((1-r.^2)/(n-2));
alpha=(1-tcdf(tvalues,n-2))*2;
sigmat=alpha<sig;
sigmatr=sigmat.*r;
if (n>m)
partvalues=abs(peparr)./sqrt((1-peparr.^2)/(n-m));
paralpha=(1-tcdf(partvalues,n-m))*2;
else
partvalues=abs(peparr)./sqrt((1-peparr.^2)/(n-2));
paralpha=(1-tcdf(partvalues,n-2))*2;
end
parsigmat=paralpha<sig;
parsigmatr=parsigmat.*peparr;
if (n<=m) threshr=rrank(parsigmatr,y1); parsigmatr=parsigmatr>=threshr; parsigmatr=parsigmatr.*peparr; end;
else
tvalues=abs(spr)./sqrt((1-spr.^2)/(n-2));
alpha=(1-tcdf(tvalues,n-2))*2;
```

```
spsigmat=alpha<sig;
spsigmatr=spsigmat.*spr;
if (n>m)
partvalues=abs(spparr)./sqrt((1-spparr.^2)/(n-m));
paralpha=(1-tcdf(partvalues,n-m))*2;
else
partvalues=abs(spparr)./sqrt((1-spparr.^2)/(n-2));
paralpha=(1-tcdf(partvalues,n-2))*2;
end;
spparsigmat=paralpha<sig;
spparsigmatr=spparsigmat.*spparr;
if (n<=m) threshr=rrank(spparsigmatr,y2); spparsigmatr=spparsigmatr>=threshr; spparsigmatr=spparsigmatr.*spparr;
end;
end;
end;
L1=(parsigmatr & spparsigmatr & sigmatr & spsigmatr);            % candidate direct interactions
L21=(parsigmatr & sigmatr); L22=(spparsigmatr & spsigmatr);     % candidate direct interactions
L3=(parsigmatr & spparsigmatr);                                  % candidate direct interactions
L41=parsigmatr; L42=spparsigmatr;                                % candidate direct interactions
L5=(sigmatr & spsigmatr);                                        % indirect/direct interactions
L61=sigmatr; L62=spsigmatr;                                      % indirect/direct interactions
for i=1:9;
switch (i)
    case 1
        mat=L1; s='L1';
    case 2
        mat=L21; s='L21';
    case 3
        mat=L22; s='L22';
    case 4
        mat=L3; s='L3';
    case 5
        mat=L41; s='L41';
    case 6
        mat=L42; s='L42';
    case 7
        mat=L5; s='L5';
    case 8
        mat=L61; s='L61';
    case 9
        mat=L62; s='L62';
end;
disp(['Interactions (the interactions found in higher levels are also listed) with statistically significance for reliability level:
' s])
[pairx,pairy]=find(mat);
temp1=pairx; temp2=pairy;
pairxs=pairx(temp1<temp2);
pairys=pairy(temp1<temp2);
InteractionPairs=[pairxs pairys]
end;
```

Two M function files, spearman.m, and rrank.m, are as follows:

```
function spearm =spearman(x,y)              %x and y: two column vectors to be tested.
if (max(size(x))~=max(size(y)))
  error('Array sizes do not match.');
end
if ((min(size(x))~=1) | (min(size(y))~=1))
  error('Both x and y are vectors');
end
n=max(size(x));
for i=1:n
rx(i)=0;ry(i)=0;xx(i)=0;yy(i)=0;
end
for j=1:n
nx=1;ny=1;
for i=1:n
if (x(i)<x(j)) nx=nx+1; end
if (y(i)<y(j)) ny=ny+1; end
end
rx(j)=nx;
ry(j)=ny;
end
for j=1:n
if (rx(j)==(n+1)) continue; end
nx=rx(j);
ntie=-1;
for i=1:n
if (rx(i)~=nx) continue; end
ntie=ntie+1;
xx(i)=rx(i);
rx(i)=0;
end
for i=1:n
if (rx(i)~=0) continue; end
xx(i)=xx(i)+(ntie*0.5);
rx(i)=n+1;
end
end
for j=1:n
if (ry(j)==(n+1)) continue; end
ny=ry(j);
ntie=-1;
for i=1:n
if (ry(i)~=ny) continue; end
ntie=ntie+1;
yy(i)=ry(i);
ry(i)=0;
end
for i=1:n
```

```
if (ry(i)~=0) continue; end
yy(i)=yy(i)+ntie*0.5;
ry(i)=n+1;
end
end
rs=0;
rs=sum((xx-yy).^2);
spearm=1-((6*rs)/(n*(n^2-1)));

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;
```

3 Application

I use a dataset on arthropod ecosystem, PH-Apr-fg (20 functional groups, 60 samples; Schoenly and Zhang, 1999; Zhang, 2011; Table 1), to calculate interactions using the algorithm above. The IDs of 20 functional groups represent:

Herbivores

(1) Pollen feeder, (2) External plant feeder, (3) Leaf roller/webber, (4) Leaf miner, (5) Gall former, (6) Mixed (combination of two or more of above).

Predators

(7) Terrestrial flyer, (8) Terrestrial crawler, walker, jumper, or hunter, (9) Neustonic (water surface) swimmer (semiaquatic), (10) Planktonic (water column) swimmer and diver, (11) Terrestrial web-builder.

Parasitoids/parasites

(12) Terrestrial blood sucker, (13) Flying adult that is searching, ovipositing, or larvipositing, (14) Idiobiont (acarine ectoparasitoid).

Detritivores

(15) Collector (filterer, suspension feeder), (16) Collector (gatherer, deposit feeder), (17) Shredder, chewer of coarse particulate Matter.

Tourists

(18) Tourist (nonpredatory species with no known functional role other than as prey in ecosystem).

Omnivores

(19) Herbivore, predator, and detritivore.

Dual insectivores

(20) Predator and parasitoid.

Table 1 Mean number of individuals per sample for each functional group.

ID of functional group	1	2	3	4	5	6	7	8	9	10
Mean number of individuals per sample	0.07	51.47	0.02	0.18	0.03	4.67	2.80	27.60	25.15	1.25
ID of functional group	11	12	13	14	15	16	17	18	19	20
Mean number of individuals per sample	2.43	0.43	5.98	0.07	1.52	2.32	5.23	2.60	0.68	0.20

Choose the significance degree p=0.01, and some of the results are as follows:

Interactions (the interactions found in higher levels are also listed) with statistically significance for reliability level: L1
InteractionPairs =

2 8

Interactions (the interactions found in higher levels are also listed) with statistically significance for reliability level: L21
InteractionPairs =

1 2
2 8
2 9
4 13
14 16
1 17
8 17
16 19

Interactions (the interactions found in higher levels are also listed) with statistically significance for reliability level: L3
InteractionPairs =

2 8

Interactions (the interactions found in higher levels are also listed) with statistically significance for reliability level: L41
InteractionPairs =

1 2
2 8

2	9
1	11
4	12
4	13
12	16
14	16
1	17
8	17
11	19
16	19

Interactions (the interactions found in higher levels are also listed) with statistically significance for reliability level: L5
InteractionPairs =

1	2
2	8
2	9
8	9
2	13
4	13
2	15
2	16
8	16
14	16
1	17
2	17
8	17
9	17
16	17
13	18
16	19
1	20

It is obvious that the most reliable interaction is (External plant feeder -- Terrestrial crawler, walker, jumper, or hunter), which represents the interaction (herbivores, predators) - the most important ecological interaction between species. The two groups are also the most abundant and predominant arthropod groups in the rice field investigated (Table 1).

4 Discussion

Our knowledge of biological networks is so limited, e.g., 80% of themolecular interactions in cells of Yeast (Yu et al., 2008) and 99.7% of human (Amaral, 2008) are still unknown. Here I have given a set of rules for finding interactions. However, these rules can be revised and improved upon the requirement of researchers.

The final results for finding, candidate direct interactions, are true direct interactions only they are

confirmed by experiments and observations. The present method provides a prompt and relatively reliable tool for primeval and batch screening of possible interactions. This will help to save time and cost for interaction finding by experiments or observations.

In a sense, the significance level is a reliability measure also. In present example, I use a significance level of p=0.001. To avoid missing candidate direct interactions as possible as, i.e., coarse screening of interactions, the significance level can be adjusted to a reasonable value, for example, p=0.05, or even p=0.1.

The validity of taxa by sample based interaction finding, as presented in this study, is dependent upon the representativeness of samples, reasonable distribution of samples over space or time, number of samples (sample size), etc. Thus how to take a set of statistically representative samples for correlation analysis is one of the keys for interaction finding.

Acknowledgment

I am thankful to the support of 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

Amaral LA. 2008. A truer measure of our ignorance. PNAS, 105(19): 6795-6796

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

Spearman C. 1904. The proof and measurement of association between two things. American Journal of Psychology, 15: 72-101

Yu H, Braun P, Yildirim MA, et al. 2008. High-quality binary protein interaction map of the yeast interactome network. Science, 322(5898): 104-110

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): 65-77

Zhang WJ, Li X. 2015a. General correlation and partial correlation analysis in finding interactions: with Spearman rank correlation and proportion correlation as correlation measures. Network Biology, 5(4): 163-168

Zhang WJ, Li X. 2015b. 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