Network Biology, 2016, 6(1): 1-11

*Article*

# A node degree dependent random perturbation method for prediction of missing links in the network

**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 25 September 2015*; *Accepted 27 October 2015*; *Published online 1 March 2016*

## Abstract

In present study, I proposed a node degree dependent random perturbation algorithm for prediction of missing links in the network. In the algorithm, I assume that a node with more existing links harbors more missing links. There are two rules. Rule 1 means that a randomly chosen node tends to connect to the node with greater degree. Rule 2 means that a link tends to be created between two nodes with greater degrees. Missing links of some tumor related networks (pathways) are predicted. The results prove that the prediction efficiency and percentage of correctly predicted links against predicted missing links with the algorithm increases as the increase of network complexity. The required number for finding true missing links in the predicted list reduces as the increase of network complexity. Prediction efficiency is complexity-depedent only. Matlab codes of the algorithm are given also. Finally, prospect of prediction for missing links is briefly reviewed. So far all prediction methods based on static topological structure only (represented by adjacency matrix) seems to be low efficient. Network evolution based, node similarity based, and sampling based (correlation based) methods are expected to be the most promising in the future.

**Keywords** missing links; network; rules; node degree; random perturbation; prediction; likelihood.

**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

Many biological networks (food webs, protein–protein interaction networks and metabolic networks, etc) are incomplete networks due to missing links. For example, 80% of the molecular interactions in cells of Yeast (Yu et al., 2008) and 99.7% interactions of human (Amaral, 2008) are unknown. An incomplete network occurs due to our limited knowledge on the network, or the network is in evolution and thus more links or even nodes are expected with time. Link (connection) prediction tries to estimate the likelihood of the existence of a link between two nodes based on observed links and (or) the attributes of nodes (Zhang, 2015d; Zhou, 2015). Link prediction can largely reduce the experimental costs for link finding. Also, link finding algorithms can be used to predict the links that may appear in the future of evolving networks (Lü and Zhou, 2011; Lü et al., 2012; Zhou, 2015). So far, numerous research on link prediction have been conducted (Clauset et al., 2008; Guimera and Sales-Pardo, 2009; Barzel and Barabási, 2013; Bastiaens et al., 2015; Lü et al., 2015; Zhang, 2015b, 2015c, 2015d, 2016b; Zhang and Li, 2015; Zhao et al., 2015; Zhou, 2015). In present study, I will propose an algorithm for prediction of missing links in the network, in which the likelihood of missing links of a node depends on the node degree.

## 2 Methods

### 2.1 Algorithm

Link prediction is closely correlated with network evolution. Following the principle of network evolution of Zhang's model (Zhang, 2016a), in present algorithm I assume that a node with more existing links harbors more missing links. It is a reasonable and practical assumption because new nodes tend to connect the nodes with more links (Barabasi and Albert, 1999; Zhang, 2012a; Zhang, 2016a).

Assume there are totally $v$ nodes in the network being predicted, and adjacency matrix of the network is $d=(d_{ij})$, $i, j=1,2,\ldots,v$, where $d_{ij}=d_{ji}$, $d_{ii}=0$, and if $d_{ij}=1$ or $d_{ji}=1$, there is a link (connection) between nodes $i$ and $j$. The adjacency matrix of the network for missing links only is $D=(D_{ij})$, $i, j=1,2,\ldots,v$. The procedures are as follows

(1) Calculate the expected missing links to be predicted, $m=m'\times per$, where $m'$ is the total links of the network, $per$ is the perturbation rate, and $per=0.2, 0.3$, etc., which represents a percentage increment of links in the network perturbation.

(2) Calculate the degree of node, $a_i(t)$, $i=1,2,\ldots,v$. The cumulative attraction strength of node 1 to node $i$ is

$$p_i(t) = \sum_{j=1}^{i} a_j(t)^{\lambda(t,a_j)} / \sum_{j=1}^{v} a_j(t)^{\lambda(t,a_j)}$$

where $\lambda$ is attraction factor, $\lambda>0$. For example, $\lambda=1.2, 1.5$, etc.

(3) Generate missing links. Let $p_0=0$, and generate two random values $w$ and $u$. For $p_0, p_1, p_2, \ldots, p_v$, one of the following two rules is used

Rule 1: if $(j-1)/v \le w \le j/v$, $p_{k-1} \le u \le p_k$, $k\ne j$, and $d_{kj}=d_{jk}=0$, let $D_{kj}=1$ and $D_{jk}=1$, i.e., there is a missing link between nodes $k$ and $j$.

Rule 2: if $p_{j-1}(t) \le w \le p_j(t)$, $p_{k-1}(t) \le u \le p_k(t)$, $k\ne j$, and $d_{kj}=d_{jk}=0$, let $D_{kj}=1$ and $D_{jk}=1$, i.e., there is a missing link between nodes $k$ and $j$.

Rule 1 means that a randomly chosen node tends to connect to the node with greater degree. Rule 2 means that a link tends to be created between two nodes with greater degrees. By doing so, a new link is found. Repeat the procedure $m$ times to produce $m$ (missing) links. By doing so, an adjacency matrix of the network for missing links only, $D=(D_{ij})$, $i, j=1,2,\ldots,v$, is generated.

(4) Return (3) to perform the next prediction, until the desired simulation times are achieved.

(5) Calculate mean number (likelihood) of predicted missing links, and rank the likelihood from greater to smaller. The first $m$ links are the predicted missing links with maximal likelihood.

The following are Matlab codes of the algorithm (linksPrediction.m)

```
%Reference: Zhang WJ. 2016. A node degree dependent random perturbation method for prediction of missing links in the network. Network Biology, 6(1): 1-11
clear
choice=input('Input the type (1 or 2) of data file of the network from which missing links are ready to be predicted (1: adjacency matrix; 2: two array): ');
```

```
disp('Adjacency matrix: d=(dij)m*m, where m is the number of nodes in the network. dij=1, if vi and vj are adjacent, and dij=0, if vi and vj are not adjacent; i, j=1,2,…, m');
disp('Two array: there are two columns, A1 and A2, in the data file; an element of A1 stores a node of a link and the corresponding element of A2 stores another node of the link. ');
if (choice==1)
adjstr=input('Input the file name of adjacency matrix from which missing links are ready to be predicted (e.g., raw.txt, raw.xls, etc. Adjacency matrix is d=(dij)m*m, where m is the number of nodes in the network. dij=1, if vi and vj are adjacent, and dij=0, if vi and vj are not adjacent; i, j=1,2,…, m: ','s');
end
if (choice==2)
adjstr=input('Input the file name of two array of the network from which missing links are ready to be predicted (e.g., raw.txt, raw.xls, etc. There are two columns, A1 and A2, in the data file; an element of A1 stores a node of a link and the corresponding element of A2 stores another node of the link: ','s');
end
rule=input('Input the rule type (1 or 2) used in the algorithm: ');
pro=input('Input perturbation rate to increase missing links of the network (e.g, 0.2, 0.3, etc.): ');
lamda=input('Attraction factor of nodes (lamda>0; e.g., 1.3, 1.5, etc.)= ');
simu=input('Input the simulation times (e.g, 100, 200, etc.): ');
if (choice==1) adjmat=load(adjstr); v=size(adjmat,2); end
if (choice==2)
twoarray=load(adjstr);
nn=size(twoarray,1);
v=max(max(twoarray));
for i=1:nn
adjmat(twoarray(i,1),twoarray(i,2))=1;
adjmat(twoarray(i,2),twoarray(i,1))=1;
end; end
degr=sum(adjmat);
m=round(sum(degr)/2*pro);
fprintf('\nAdjacency matrix of the original network\n')
disp([adjmat])
fprintf('\nNode degrees of adjacency matrix of the original network\n')
disp([degr])
fprintf(['\nMean of node degrees of the original network: ' num2str(mean(degr)) '\n\n'])
cnow=(sum(degr)/2)/((v^2-v)/2);
fprintf(['\nConnectance=' num2str(cnow) '\n'])
summ=sum(degr);
summa=sum(degr.*(degr-1));
h=v*summa/(summ*(summ-1));
fprintf(['\nAggregation index (AI) of node degrees=' num2str(h) '\n'])
cv=(std(degr))^2/mean(degr);
fprintf(['\nCoefficient of variation (CV) of node degrees=' num2str(cv) '\n'])
summ=v*(v-1)/2;
su=zeros(summ,2*simu);
prop=zeros(1,v);
```

```
proptot=zeros(v);
degrr=degr.^lamda;
prop(1)=degrr(1)/sum(degrr);
for i=2:v;
prop(i)=prop(i-1)+degrr(i)/sum(degrr);
end
for siml=1:simu
adj=zeros(v);
temp=zeros(m,2);
mm=1;
while (v>0)
rep=0;
while (v>0)
propp=prop;
if ((rep==0) & (rule==1))
for i=1:v;
propp(i)=i/v;
end; end
ran=rand();
for j=1:v
if (j==1) st=0; end
if (j>=2) st=propp(j-1); end
if ((ran>=st) & (ran<propp(j))) rep=rep+1; id(rep)=j; break; end
end
if ((rep>=2) & (id(rep)~=id(1)))
tab=0;
for i=1:mm
if (((id(1)==temp(i,1)) & (id(rep)==temp(i,2))) | ((id(rep)==temp(i,1)) & (id(1)==temp(i,2)))) tab=1; break; end
end
if (tab==1) continue; end;
temp(mm,1)=id(1); temp(mm,2)=id(rep);
break;
end; end
if (adjmat(id(1),id(rep))==0) adj(id(1),id(rep))=1; adj(id(rep),id(1))=1; mm=mm+1; end;
if (mm==m+1) break; end;
end
fprintf(['Simulation ' num2str(siml)])
fprintf('\n\nAdjacency matrix for predicted links only\n')
disp([adj])
[pairx,pairy]=find(adj);
temp1=pairx; temp2=pairy;
pairxs=pairx(temp1<temp2);
pairys=pairy(temp1<temp2);
ConnectionPairs=[pairxs pairys];
dm=size(ConnectionPairs,1);
```

```
su(:,siml*2-1)=[pairxs;zeros(summ-dm,1)]; su(:,siml*2)=[pairys;zeros(summ-dm,1)];
disp('Predicted links')
disp([ConnectionPairs])
end
disp('--------------------------------Summary--------------------------------')
disp(['There are totally ' num2str(sum(degr)/2) ' links in the original network'])
disp(['You wish to predict ' num2str(m) ' missing links in the original network'])
fprintf('\n');
proptot=zeros(v);
for i=1:v-1
for j=i+1:v
for k=1:simu
for l=1:v*(v-1)/2
if ((su(l,k*2-1)==i) & (su(l,k*2)==j)) proptot(i,j)=proptot(i,j)+1; proptot(j,i)=proptot(i,j); break; end
end; end; end; end
disp('Likelihood (mean number) of predicted links: ')
disp('      Node          Node          Likelihood')
s=0;
for j=1:v
for i=1:v
if (proptot(i,j)~=0) s=s+1;pairvalue(s)=proptot(i,j)/simu; end;
end; end
[pairx,pairy]=find(proptot);
result=[pairx pairy pairvalue'];
results(1,1)=result(1,1); results(1,2)=result(1,2); results(1,3)=result(1,3);
su=1;
for i=2:s
lab=0;
for j=1:i-1
if ((result(j,2)==result(i,1)) & (result(j,1)==result(i,2))) lab=1; break; end;
end
if (lab==0) su=su+1;results(su,1)=result(i,1); results(su,2)=result(i,2); results(su,3)=result(i,3); end
end
ires=sortrows(results,-3);
disp([ires])
```

## 2.2 Validation

In present study, I used the data of tumor related networks (pathways) (ABCAM, 2012; Huang and Zhang, 2012; Li and Zhang, 2013; Pathway Central, 2012; See supplementary material for adjacency matrices). These networks are complete. For each network, some links are removed following reverse process of the algorithm above and then predicted. The simulation times are set to be 100. The perturbation rate is *per*=~0.25. Attraction factor $\lambda$=1.5.

## 3 Results

### 3.1 Rule 1

Some of the summarized results for link prediction of tumor related networks (the pathways Ras, p53, Akt, HGF, JNK, PPAR, TGF-β, and TNF) are listed in Table 1 and 2, and the percentages of correctly predicted links with randomization method are given also. Here, the percentage of correctly predicted links against number of missing links (%) = correctly predicted links / number of missing links ×100, and the percentage of correctly predicted links against predicted missing links (%) = correctly predicted links / total of predicted missing links ×100, connectance = number of observed links / number of possible maximum number of links.

**Table 1** Link prediction of Ras, p53, and Akt networks with Rule 1 (*per*=~0.25, $\lambda$=1.5, 100 simulations). The listed links are true links missed in the data used for predicting.

| Ras | | | | p53 | | | | Akt | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Rank | Node | Node | Likelihood | Rank | Node | Node | Likelihood | Rank | Node | Node | Likelihood |
| 28 | 9 | 5 | 0.04 | 82 | 47 | 32 | 0.04 | 465 | 35 | 31 | 0.01 |
| 34 | 28 | 5 | 0.04 | 138 | 47 | 33 | 0.03 | 151 | 50 | 12 | 0.02 |
| 58 | 22 | 5 | 0.03 | 140 | 47 | 36 | 0.03 | 2 | 51 | 15 | 0.18 |
| 137 | 10 | 5 | 0.02 | 88 | 48 | 47 | 0.04 | 1 | 51 | 16 | 0.2 |
| 140 | 25 | 5 | 0.02 | 11 | 52 | 4 | 0.07 | 26 | 51 | 24 | 0.1 |
| 230 | 31 | 28 | 0.02 | 61 | 52 | 9 | 0.04 | 17 | 51 | 28 | 0.12 |
| 392 | 35 | 34 | 0.01 | 4 | 52 | 10 | 0.09 | 28 | 51 | 31 | 0.1 |
| | | | | 269 | 52 | 30 | 0.02 | 36 | 51 | 38 | 0.07 |
| | | | | 18 | 52 | 48 | 0.07 | 10 | 51 | 39 | 0.14 |
| | | | | 19 | 52 | 51 | 0.07 | 31 | 51 | 41 | 0.09 |
| | | | | | | | | 7 | 51 | 42 | 0.15 |
| | | | | | | | | 20 | 52 | 51 | 0.12 |

According to Table 1 and 2, the regression relationships between aggregation index ($u$), coefficient of variation ($w$) (Zhang and Zhan, 2011; Zhang, 2012a), and prediction efficiency ($z=x/y$, where $x$ is the percentages of correctly predicted links, and $y$ is the averaged ranks before which all missing links fall in the list of predicted links), the percentage (%) of correctly predicted links against predicted missing links ($q$), and the rate of the averaged rank before which all missing links fall in the list of predicted links *vs*. total number of predicted missing links ($f$) are as follows

Algorithm prediction:

$z=0.320+0.344u \quad r^2=0.318,\ p=0.019<0.05,\ n=17$

$z=0.465+0.192w \quad r^2=0.323,\ p=0.017<0.05,\ n=17$

$q=1.349+0.243u \quad r^2=0.106,\ p=0.203,\ n=17$

$q=1.427+0.154u \quad r^2=0.139,\ p=0.141,\ n=17$

$f=0.438-0.125u \quad r^2=0.306,\ p=0.021<0.05,\ n=17$

$f=0.389-0.073w \quad r^2=0.341,\ p=0.014<0.05,\ n=17$

Randomization prediction:

$z=0.485-0.106u \quad r^2=0.149,\ p=0.125,\ n=17$

$z=0.445-0.063w \quad r^2=0.171,\ p=0.099<0.1,\ n=17$

$q=1.615-0.349u \quad r^2=0.259, p=0.038<0.05, n=17$

$q=1.451-0.182u \quad r^2=0.229, p=0.051<0.01, n=17$

$f=0.476-0.088u \quad r^2=0.156, p=0.117, n=17$

$f=0.436-0.046w \quad r^2=0.142, p=0.136, n=17$

Thus prediction efficiency and the percentage of correctly predicted links against predicted missing links with the algorithm increases as the increase of network complexity. Generally, the rate of averaged rank of true missing links in the list of predicted missing links declines as the network complexity, which means the required number for checking true missing links in the predicted list reduces as the increase of network complexity.

Compared to the prediction of randomization method, in general, the results of the algorithm are effective, i.e., the present algorithm is effective in predicting missing links of biological networks (Table 1, 2).

Both mean of node degrees and connectance have not significant relationships with prediction efficiency. Thus prediction efficiency is complexity-depedent only.

**Table 2** Link prediction of some tumor related networks of missing links with Rule 1 (*per*=~0.25, λ=1.5).

| | PPAR | TGF-β | TNF | STAT3 | mTOR | Ras | EGF | PTEN | JAK-STAT |
|---|---|---|---|---|---|---|---|---|---|
| Mean of node degrees | 1.85 | 1.79 | 2.06 | 1.75 | 1.83 | 1.71 | 1.96 | 2.06 | 2.09 |
| Connectance | 0.07 | 0.05 | 0.07 | 0.08 | 0.04 | 0.05 | 0.04 | 0.06 | 0.05 |
| Possible maximum number of candidate links | 326 | 669 | 433 | 255 | 993 | 565 | 1431 | 494 | 858 |
| Aggregation Index (Zhang and Zhan, 2011; Zhang, 2012a ) | 0.68 | 0.78 | 0.85 | 0.72 | 0.75 | 0.75 | 0.73 | 0.91 | 0.91 |
| Coefficient of variation (Zhang and Zhan, 2011; Zhang, 2012a) | 0.40 | 0.61 | 0.68 | 0.51 | 0.54 | 0.57 | 0.47 | 0.82 | 0.81 |
| **Percentage (%) of correctly predicted links against true missing links with the algorithm (*x*)** | **83.3** | **75.0** | **87.5** | **100** | **80.0** | **87.5** | **84.6** | **75.0** | **45.5** |
| **Percentage (%) of correctly predicted links against predicted missing links with the algorithm** | **1.9** | **1.3** | **2.0** | **2.6** | **1.4** | 1.8 | 1.3 | 1.7 | **0.9** |
| Number of missing links | 6 | 8 | 8 | 5 | 10 | 8 | 13 | 8 | 12 |
| Total number of predicted links with 100 simulations | 257 | 448 | 346 | 195 | 575 | 392 | 823 | 348 | 545 |
| **The averaged rank before which all missing links fall in the list of predicted links (*y*)** | **115** | **190** | **179** | **114** | **202** | **127** | **432** | **47** | **99** |
| **Prediction efficiency (*x*/*y*)** | **0.7243** | **0.3947** | **0.4888** | **0.8772** | **0.396** | **0.689** | **0.1958** | **1.5957** | **0.4596** |
| **Percentage (%) of correctly predicted links against true missing links with randomization method (*x*)** | **100** | **75** | **87.5** | **60.0** | **70.0** | **37.5** | **61.5** | **100** | **45.5** |
| Percentage (%) of correctly predicted links against predicted missing links with randomization method | 2.2 | 1.3 | 1.9 | 1.4 | 1.1 | 0.7 | 0.9 | 2.0 | 0.8 |
| Total number of predicted links with 100 simulations | 270 | 466 | 375 | 217 | 651 | 424 | 853 | 398 | 617 |
| **The averaged rank before which all missing links fall in the list of predicted links (*y*)** | **120** | **148** | **175** | **84** | **338** | **103** | **239** | **302** | **102** |
| **Prediction efficiency (*x*/*y*)** | **0.8333** | **0.5068** | **0.5** | **0.7143** | **0.2071** | **0.3641** | **0.2573** | **0.3311** | **0.4461** |

## 3.2 Rule 2

In the step (3) of the algorithm, I use the Rule 2 for prediction. The results for some pathways are listed in Table 3. Compared to the Rule 1, the percentages (%) of correctly predicted links with the algorithm calculated from the Rule 2 are overall smaller. However, the prediction efficiency of Rule 2 is generally higher. The major regression relationships and conclusions are similar to Rule 1. Moreover, the prediction efficiency of the algorithm increases dramatically as the network complexity.

**Table 2** (continue) Link prediction of some tumor related networks of missing links with Rule 1 (*per*=~0.25, $\lambda$=1.5).

| | p53 | Akt | HGF | JNK | PI3K | MARK | FAS | ERK |
|---|---|---|---|---|---|---|---|---|
| Mean of node degrees | 1.96 | 1.69 | 1.67 | 2.67 | 2.25 | 2.14 | 1.88 | 2.27 |
| Connectance | 0.04 | 0.03 | 0.05 | 0.06 | 0.04 | 0.04 | 0.04 | 0.04 |
| Possible maximum number of candidate links | 1275 | 1604 | 600 | 1064 | 1532 | 1591 | 1277 | 1702 |
| Aggregation Index (Zhang and Zhan, 2011; ; Zhang, 2012a) | 1.50 | 3.59 | 0.96 | 1.72 | 0.97 | 1.22 | 1.17 | 1.41 |
| Coefficient of variation (Zhang and Zhan, 2011; Zhang, 2012a) | 1.99 | 5.42 | 0.93 | 2.96 | 0.93 | 1.46 | 1.32 | 1.93 |
| **Percentage (%) of correctly predicted links against true missing links with the algorithm (*x*)** | **76.9** | **100** | **57.1** | **100** | **68.8** | **53.3** | **75.0** | **94.1** |
| **Percentage (%) of correctly predicted links against predicted missing links with the algorithm** | 1.6 | 2.2 | 1.1 | 2.6 | 1.2 | 0.9 | 1.4 | 1.8 |
| Number of missing links | 13 | 12 | 7 | 16 | 16 | 14 | 12 | 17 |
| Total number of predicted links with 100 simulations | 642 | 542 | 354 | 612 | 899 | 819 | 640 | 883 |
| **The averaged rank before which all missing links fall in the list of predicted links (*y*)** | **64** | **66** | **102** | **93** | **240** | **202** | **80** | **179** |
| **Prediction efficiency (*x*/*y*)** | **1.2016** | **1.5152** | **0.5598** | **1.0753** | **0.2867** | **0.2639** | **0.9375** | **0.5257** |
| **Percentage (%) of correctly predicted links against true missing links with randomization method (*x*)** | **61.5** | **25.0** | **71.4** | **75.0** | **68.8** | **66.7** | **75.0** | **76.5** |
| Percentage (%) of correctly predicted links against predicted missing links with randomization method | 0.9 | 0.3 | 1.2 | 1.4 | 1.1 | 1.0 | 1.2 | 1.2 |
| Total number of predicted links with 100 simulations | 823 | 862 | 423 | 839 | 990 | 974 | 770 | 1073 |
| **The averaged rank before which all missing links fall in the list of predicted links (*y*)** | **343** | **106** | **219** | **296** | **409** | **262** | **177** | **505** |
| **Prediction efficiency (*x*/*y*)** | **0.1793** | **0.2358** | **0.326** | **0.2534** | **0.1682** | **0.2546** | **0.4237** | **0.1515** |

**Table 3** Link prediction of some tumor related networks (pathways) of missing links with Rule 2 (*per*=~0.25, $\lambda$=1.5).

| | Ras | p53 | Akt | HGF | JNK | PPAR | TGF-β | TNF |
|---|---|---|---|---|---|---|---|---|
| **Percentage (%) of correctly predicted links with the algorithm (*x*)** | **62.5** | **92.3** | **50.0** | **57.1** | **75.0** | **33.3** | **37.5** | **62.5** |
| Percentage (%) of correctly predicted links against predicted missing links with the algorithm | 0.7 | 0.9 | 0.2 | 1.5 | 1.2 | 1.8 | 1.5 | 1.7 |
| Total number of predicted links with 100 simulations | 314 | 388 | 301 | 300 | 404 | 221 | 304 | 291 |
| **The averaged rank before which all missing links fall in the list of predicted links (*y*)** | **74** | **92** | **6** | **78** | **81** | **41** | **63** | **95** |
| **Prediction efficiency (*x*/*y*)** | **0.8446** | **1.0033** | **8.3333** | **0.7321** | **0.9259** | **0.8122** | **0.5952** | **0.6579** |
| **Percentage (%) of correctly predicted links against number of missing links with random network (*x*)** | **37.5** | **53.9** | **16.7** | **85.7** | **62.5** | **83.3** | **87.5** | **75.0** |
| Percentage (%) of correctly predicted links against predicted missing links with random network | 1.6 | 3.1 | 2.0 | 1.3 | 2.9 | 0.9 | 0.9 | 1.7 |
| Total number of predicted links with 100 simulations | 411 | 823 | 851 | 412 | 838 | 277 | 478 | 359 |
| **The averaged rank before which all missing links fall in the list of predicted links (*y*)** | **77** | **325** | **23** | **213** | **246** | **55** | **184** | **111** |
| **Prediction efficiency (*x*/*y*)** | **0.487** | **0.1658** | **0.7261** | **0.4023** | **0.2541** | **1.5145** | **0.4755** | **0.6757** |

## 4 Discussion

As stated above, random prediction is overall effective for the random networks only. However, in practical applications, most networks are complex networks. Thus the algorithm is effective in predicting missing links in most cases. The prediction efficiency of the algorithm increases as the increase of network complexity. Therefore, the algorithm is more efficient for the networks of higher complexity.

The changes of $\lambda$ can reflect various effects of the node degree on connection mechanism. The larger $\lambda$ will lead to find more missing links of the nodes with greater node degree. $\lambda \rightarrow 0$ means a trend to random prediction. How to fix a suitable value of $\lambda$, is specific to practical problems.

Lü et al. (2015) proposed the structural perturbation method (SPM) to predict missing links and argued that its prediction ability was stronger than previous methods. However, I affim their method does not hold due to the following reasons: (1) Mechanically, the structural perturbation method can only be used to analyze structural stability of dynamic systems. The static structure of a network, expressed by an adjacent matrix, is the topological structure, which cannot represent the dynamic charicteristics of the network evolution. Pediction of missing links should be conducted on the basis of mechanism of network evolution (dynamics). Without loss of generality, network evolution may be approximated with a group of linear differential equations (Zhang, 2015a). And the structural stability of the network was determined by the eigenvalues but not eigenvectors of system matrix. Even so, the structural perturbation method for determining the variables with least impact on structural stability should only be used around the equilibrium states of the system rather than the states far away the equilibrium. (2) During the evolution of a network, the generated links with most likehood are not necessarily those links that minimaly perturb the topological structure of the network. On the premise of not destroying the structural stability of the system and no other limitations, any links will prepare to be created. A most occurred case is that two nodes with most similarity will firstly connect to each other. (3) Utilization of missing links in the prediction model to predict missing links, as done by Lü et al. (2015), is somewhat similar to model fitting but not prediction. In this case, the stronger “prediction” ability (precisely, fitting ability) is surely expected.

So far all prediction methods based on static topological structure only (represented by adjacency matrix) seems to be low efficient. Network evolution based (Zhang, 2012a, 2012c, 2015a, 2016a, 2016b), node similarity based (Zhang, 2015d), and sampling based (correlation based; Zhang, 2007, 2011, 2012b, 2013, 2015b; Zhang and Li, 2015) methods are expected to be the most promising in the future.

## 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

ABCAM. 2012. http://www.abcam.com/index.html?pageconfig=productmap&cl=2282

Amaral LAN. 2008. A truer measure of our ignorance. Proceedings of the National Academy of Sciences of USA, 105: 6795-6796

Barabasi AL, Albert R. 1999. Emergence of scaling in random networks. Science, 286(5439): 509

Barzel B, Barabási AL. 2013. Network link prediction by global silencing of indirect correlations. Nature Biotechnology, 31: 720-725

Bastiaens P, Birtwistle MR, Blüthgen N, et al. 2015. Silence on the relevant literature and errors in implementation. Nature Biotechnology, 33: 336-339

Clauset A, Moore C, Newman MEJ. 2008. Hierarchical structure and the prediction of missing links in networks. Nature, 453: 98-101

Guimera R, Sales-Pardo M. 2009. Missing and spurious interactions and the reconstruction of complex networks. Proceedings of the National Academy of Sciences of USA, 106: 22073-22078

Huang JQ, Zhang WJ. 2012. Analysis on degree distribution of tumor signaling networks. Network Biology, 2(3): 95-109

Li JR, Zhang WJ. 2013. Identification of crucial metabolites/reactions in tumor signaling networks. Network Biology, 3(4): 121-132

Lü LY, Medo M, Yeung CH, et al. 2012. Recommender systems. Physics Reports, 519: 1-49

Lü LY, Pan LM, Zhou T, et al. 2015. Toward link predictability of complex networks. Proceedings of the National Academy of Sciences of USA, 112: 2325-2330

Lü LY, Zhou T. 2011. Link prediction in complex networks: A survey. Physica A, 390: 1150-1170

Pathway Central. 2012. SABiosciences. http://www.sabiosciences.com/pathwaycentral.php

Yu HY, Braun P, Yildirim MA, et al. 2008. High-quality binary protein interaction map of the yeast interactome network. Science, 322: 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. 2012c. Modeling community succession and assembly: A novel method for network evolution. Network Biology, 2(2): 69-78

Zhang WJ. 2013. Construction of Statistic Network from Field Sampling. In: Network Biology: Theories, Methods and Applications (WenJun Zhang, ed). 69-80, Nova Science Publishers, New York, USA

Zhang WJ, 2015a. A generalized network evolution model and self-organization theory on community assembly. Selforganizology, 2(3): 55-64

Zhang WJ. 2015b. A hierarchical method for finding interactions: Jointly using linear correlation and rank correlation analysis. Network Biology, 5(4): 137-145

Zhang WJ. 2015c. Calculation and statistic test of partial correlation of general correlation measures. Selforganizology, 2(4): 65-77

Zhang WJ. 2015d. Prediction of missing connections in the network: A node-similarity based algorithm. Selforganizology, 2(4): 91-101

Zhang WJ. 2016a. A random network based, node attraction facilitated network evolution method. Selforganizology, 3(1): 1-9

Zhang WJ. 2016b. Selforganizology: The Science of Self-Organization. World Scientific, Singapore

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, Liu GH. 2012. Creating real network with expected degree distribution: A statistical simulation. Network Biology, 2(3): 110-117

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): 159-170

Zhao J, Miao LL, Yang Y, et al. 2015. Prediction of links and weights in networks by reliable routes. Scientific Reports, 5: 12261

Zhou T. 2015. Why link prediction? http://blog.sciencenet.cn/blog-3075-912975.html. Accessed on Aug 14, 2015