Article

Prediction of missing connections in the network: A node-similarity based algorithm

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 18 August 2015; *Accepted 21 September 2015*; *Published online 1 December 2015*

Abstract
In present study, I proposed a node-similarity based algorithm for prediction of missing connections in the network. In this algorithm, whether a node v_k can connect to v_i or not, depending on the similarity between v_k and v_i, the similarities between v_i and its adjacent nodes, the similarities between v_k and the adjacent nodes of v_i, and the degree of node v_i, and vice versa. Pearson correlation measure, cosine measure, and (negative) Euclidean distance measure (the three measures are for interval attributes), contingency correlation measure (for nominal attributes), and Jaccard coefficient measure (for binary attributes) were used as the between-node similarity. Two application examples showed a better prediction of the algorithm (approximately 60% of missing connections are successfully predicted). Matlab codes of the algorithm were provided.

Keywords network; connections; prediction; node similarity; algorithm.

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

Connection prediction aims to estimate the likelihood of the existence of a connection between two nodes, based on observed connections and the attributes of nodes (Zhou, 2015). Many biological networks, such as food webs, protein–protein interaction networks and metabolic networks, are incomplete networks due to missing connections. For example, 80% of the molecular interactions in cells of Yeast (Yu et al., 2008) and 99.7% of human (Amaral, 2008) are still unknown. An incomplete network occurs due to our limited knowledge on a complete network, or the network is in evolution and thus more connections or even nodes are expected with time. Connection prediction can considerably reduce the experimental costs for connection finding. Moreover, connection prediction algorithms can be used to predict the connections that may appear in the future of evolving networks (Lü and Zhou, 2011; Lü et al., 2012; Zhou, 2015). So far, connection prediction has attracted wide attention. Numerous papers on this topic have been published (Clauset et al., 2008; Guimera R, Sales-Pardo, 2009; Barzel and Barabási, 2013; Bastiaens et al., 2015; Lü et al., 2015; Zhang, 2015a, 2015b; Zhang and Li, 2015; Zhao et al., 2015; Zhou, 2015).

In present study, I will propose a node-similarity based algorithm for prediction of missing connections in a network. Matlab codes of the algorithm will be presented for further use.

2 Algorithm

Suppose there is an incomplete network, *X*, with *m* nodes (Zhang, 2012a), its adjacency matrix is $d=(d_{ij})_{m\times m}$. $d_{ij}=1$, if two nodes v_i and v_j are adjacent, and $d_{ij}=0$, if v_i and v_j are not adjacent; $i, j=1,2,\ldots, m$. Adjacency matrix *d* is a symmetric matrix, i.e., $d=d'$. Known *n* attributes for *m* nodes. The raw data matrix is $a=(a_{ij})_{m\times n}$. Pearson correlation measure, cosine measure, and (negative) Euclidean distance measure (the three measures are for interval attributes), contingency correlation measure (for nominal (1, 2, 3…) attributes), and Jaccard coefficient measure (for binary (0, 1) attributes) can be as the between-node similarity (Zhang, 2016).

Pearson correlation measure is (Zhang, 2011; Zhang et al., 2014; Zhang, 2012a, b; Zhang and Li, 2015)

$$r_{ij}= \sum_{k=1}^{n} ((a_{ik} - a_{ib})(a_{jk}- a_{jb}))/(\sum_{k=1}^{n} (a_{ik} - a_{ib})^2 \sum_{k=1}^{n} (a_{jk} - a_{jb})^2)^{1/2}$$
$$i, j=1, 2,\cdots,m$$

where $-1\le r_{ij}\le 1$, $a_{ib}=\sum_{k=1}^{n} a_{ik}/n$, $a_{jb}=\sum_{k=1}^{n} a_{jk}/n$, $i, j=1, 2,\cdots,m$.

Cosine measure is (Zhang, 2007; Zhang, 2012a)

$$r_{ij}= \sum_{k=1}^{n} a_{ik}\, a_{jk}/(\sum_{k=1}^{n}a_{ik}^2 \sum_{k=1}^{n}a_{jk}^2)^{1/2}$$
$$i, j=1, 2,\cdots,m$$

Euclidean distance measure is (Zhang, 2007, 2012a)

$$d_{ij}= (\sum_{k=1}^{n} (a_{ik} - a_{jk})^2)^{1/2}$$

Thus its negative value is used as the similarity measure

$$r_{ij}= -d_{ij}$$

Contingency correlation measure is (Zhang, 2007, 2012a; Zhang et al., 2014):

$$r_{ij}=2(h/(s\,(p-1)))^{1/2}-1 \quad i, j=1, 2,\cdots,m$$

where $-1\le r_{ij}\le 1$, and

$$h= s_{..}(\sum^{p}_{i=1}\sum^{p}_{j=1}s_{ij}^2/(s_{i.}\, s_{.j})-1)$$
$$s_{.}=\sum^{p}_{i=1} s_{i.}\,, \quad s_{i.} =\sum^{p}_{j=1} s_{ij}\,, \quad n_{.j} =\sum^{p}_{i=1} s_{ij}$$

where there are *p* available nominal values, i.e., $t_1, t_2,\ldots, t_p$, for attributes *i* and *j*, s_{kl} is the number of attributes of node *i* takes value t_k and node *j* takes value t_l, $k, l= 1, 2, \ldots, p$.

Jaccard coefficient measure is (Zhang, 2015b)

$$r_{ij}=(e-(c+b))/(e+c+b) \quad i, j=1, 2,\cdots,m$$

where $-1 \leq r_{ij} \leq 1$, c is the number of node pairs of 1 for attribute i but not for j; b is the number of node pairs of 1 for attribute j but not for i; e is the number of node pairs of 1 for both attribute i and attribute j.

Between-node similarity matrix, $r=(r_{ij})_{m\times m}$, is a symmetric matrix, i.e., $r=r'$. In this algorithm, whether a node v_k can connect to v_i or not, depending on the similarity between v_k and v_i, the similarities between v_i and its adjacent nodes, the similarities between v_k and the adjacent nodes of v_i, and the degree of node v_i, and vice versa. The procedures of the algorithm for prediction of missing connections are as follows.

(1) For each node v_i, i=1, 2,⋯,m, and $\forall v_j \in S_i=\{v_k | d_{ik}=1\}$, calculate mean similarity between v_i and $\forall v_j \in S_i$, and mean similarity of $\forall v_j \in S_i$,

$$i\_\text{mean}=\text{mean } r_{ij} \qquad v_j \in S_i$$
$$i\_adj\_\text{mean}=\text{mean } r_{kj} \qquad v_k \in S_i,\ v_j \in S_i$$

(2) For a node v_i, i=1, 2,⋯,m, the nodes $\forall v_j \in S_i$, and the node $v_k \notin S_i$, k=1, 2,⋯,m. First, calculate the mean similarity of between v_k and $\forall v_j \in S_i$

$$k\_i\_adj\_\text{mean}=\text{mean } r_{kj} \qquad v_j \in S_i$$

then calculate the similarity win

$$z_{ki}=\alpha(r_{ki} - i\_\text{mean})+(1-\alpha)(k\_i\_adj\_\text{mean} - i\_adj\_\text{mean})$$

where $v_k \neq v_i$, $d_{ki}=0$, and α is the importance weight of node v_i against its adjacent node set S_i in determining whether the node v_k can connect to the node v_i or not, $0 \leq \alpha \leq 1$. Reverse v_k and v_i, and repeat the step (1) and (2), calculate z_{ik}.

(3) Calculate $z^{ik}=z_{ki}\, n_i/(n_k+n_i)+z_{ik}n_k/(n_k+n_i)$, where n_i is the degree of node v_i, i, k=1, 2,⋯,m; $v_k \neq v_i$, $d_{ki}=0$. The weights, $n_i/(n_k+n_i)$ and $n_k/(n_k+n_i)$, are given because the nodes of greater degree are generally more important (Barabasi and Albert, 1999; Zhang and Zhan, 2011; Huang and Zhang, 2012; Zhang, 2012c; Li and Zhang, 2013), and the calculation results on the nodes of greater degree are more statistically confident.

Finally, for $z^{ik} \geq 0$, calculate $y^{ik}=z^{ik}/2$, to achieve the averaged similarity win, which represents an averaged similarity win of a predicted missing connections against existing connections of the two nodes to be connected.

(4) Rank predicted node pairs from the larger y^{ik} to small ones. The predicted connections with the larger y^{ik} have higher confidence degree.

(5) Once some of the predicted connections are confirmed by observations, the adjacency matrix $d=(d_{ij})$, can be revised; return step (1) to start new round of prediction.

The following are Matlab codes of the algorithm

```
%Reference: Zhang WJ. 2015. Prediction of missing connections in the network: A node-similarity based algorithm.
%Selforganizology, 2(4): 91-101
raw=input('Input the file name of raw data (e.g., raw.txt, raw.xls, etc. The matrix is z=(zij)m×n, where m is total number of nodes, n is the number of attributes ): ','s');
adj=input('Input the file name of adjacency matrix or its two-array form (e.g., adj.txt, adj.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; two array form of adjacency matrix, the 1st column is from nodes and 2nd column is to nodes.): ','s');
choice=input('Input a number to choose similarity measure (1: Pearson linear correlation; 2: Cosine measure; 3: (Negative)
Euclidean distance; 4: Contingency correlation; 5: Jaccard coefficient): ');
alpha=input('Input a weight between 0 and 1 for importance of a node to be connected to against its adjacent nodes (e.g., 0.5, etc. weight=1, means absolute importance of a node and no function of its adjacent nodes): ');
raw=load(raw); m=size(raw,1); n=size(raw,2);
adj=load(adj);
if (size(adj,2)==2)
nn=size(adj,1);
adjj=zeros(m);
for i=1:nn
adjj(adj(i,1),adj(i,2))=1;
adjj(adj(i,2),adj(i,1))=1;
end
adj=adjj;
end
r=zeros(m);
for i=1:m-1
for j=i+1:m
ix=raw(i,:); jx=raw(j,:);
if (choice==1)
str='Pearson correlation';
ixbar=mean(ix);
jxbar=mean(jx);
aa=sum((ix-ixbar).*(jx-jxbar));
bb=sum((ix-ixbar).^2);
cc=sum((jx-jxbar).^2);
r(i,j)=aa/sqrt(bb*cc);
end
if (choice==2)
str='Cosine measure';
aa=sum(ix.*jx);
bb=sum(ix.^2);
cc=sum(jx.^2);
r(i,j)=aa/sqrt(bb*cc);
end
if (choice==3)
str='(Negative) Euclidean distance';
r(i,j)=-sqrt(sum((ix-jx).^2));
end
if (choice==4)
str='Contingency correlation';
xx=[ix;jx];
pn=1;
tt(1)=xx(1);
for kk=1:max(size(xx))
jj=0;
```

```
for ii=1:pn
if (xx(kk)~=tt(ii)) jj=jj+1; end;
end
if (jj==pn) pn=pn+1;tt(pn)=xx(kk); end;
end
for kk=1:pn
for jj=1:pn
temp(kk,jj)=0;
for ii=1:max(size(ix))
if ((ix(ii)==tt(kk)) & (jx(ii)==tt(jj))) temp(kk,jj)=temp(kk,jj)+1; end; end
end; end
for kk=1:pn
pp=0;
for jj=1:pn pp=pp+temp(kk,jj); end
ni(kk)=pp;
end
for kk=1:pn
pp=0;
for jj=1:pn pp=pp+temp(jj,kk); end
nj(kk)=pp;
end
summ=0;
for kk=1:pn
summ=summ+ni(kk);
end;
xsquare=0;
for kk=1:pn
for jj=1:pn
if (ni(kk)==0 | nj(jj)==0) continue; end
xsquare=xsquare+temp(kk,jj)*temp(kk,jj)/(ni(kk)*nj(jj));
end; end
xsquare=summ*(xsquare-1);
r(i,j)=2*sqrt(xsquare/(summ*(pn-1)))-1;
end
if (choice==5)
str='Jaccard coefficient';
bb=sum((ix==0) & (jx~=0));
cc=sum((ix~=0) & (jx==0));
dd=sum((ix~=0) & (jx~=0));
r(i,j)=(dd-(cc+bb))/(dd+cc+bb);
end
r(j,i)=r(i,j);
end; end
fprintf('\nPredicted potential connections, similarity win, and similarity\n')
disp('     Node         Node       Similarity win      Similarity')
```

```
nn=0;
res=zeros(m*m,3);
ilinkmean=zeros(1,m); inlinkmean=zeros(1,m);
num=zeros(1,m); nu=zeros(1,m);
for i=1:m
nu(i)=0;
ilinkmean(i)=0;
for j=1:m
if (i==j) continue; end;
if (adj(i,j)~=0)
nu(i)=nu(i)+1;
num(nu(i))=j;
ilinkmean(i)=ilinkmean(i)+r(i,j);
end; end;
ilinkmean(i)=ilinkmean(i)/nu(i);
inlinkmean(i)=0;
if (nu(i)>1)
for j=1:nu(i)-1
for k=j+1:nu(i)
if (k==j) continue; end
inlinkmean(i)=inlinkmean(i)+r(num(j),num(k));
end; end;
inlinkmean(i)=inlinkmean(i)/((nu(i)^2-nu(i))/2); end
if (nu(i)==1)
for j=1:m
if (adj(i,j)~=0) inlinkmean(i)=r(i,j); end
end; end
jinlinkmean=zeros(1,m);
for j=1:m
if ((j==i) | (sum(j==num)==1) | (adj(i,j)~=0)) continue; end
jinlinkmean(j)=0;
for k=1:nu(i)
jinlinkmean(j)=jinlinkmean(j)+r(j,num(k));
end
jinlinkmean(j)=jinlinkmean(j)/nu(i);
z=alpha*(r(i,j)-ilinkmean(j))+(1-alpha)*(jinlinkmean(j)-inlinkmean(j));
nn=nn+1; res(nn,1)=i; res(nn,2)=j; res(nn,3)=z;
end; end
ress=zeros(m*m,4);
mm=0;
for i=1:m-1
for j=i+1:m
for k=1:nn
if ((res(k,1)==i) & (res(k,2)==j)) mm=mm+1; ress(mm,4)=r(i,j); ress(mm,1)=i; ress(mm,2)=j;
ress(mm,3)=ress(mm,3)+res(k,3)*nu(i)/(nu(i)+nu(j)); end;
```

```
if ((res(k,1)==j) & (res(k,2)==i)) ress(mm,3)=ress(mm,3)+res(k,3)*nu(j)/(nu(i)+nu(j)); end
end; end; end
ress(:,3)=round(ress(:,3)/2*10000)/10000;
ress(:,4)=round(ress(:,4)*10000)/10000;
iress=zeros(mm,4);
id=0;
for i=1:mm
if (ress(i,3)>=0) id=id+1; iress(id,:)=ress(i,:); end
end
ires=sortrows(iress(1:id,:),-3);
disp([ires])
```

Table 1 Predictions of connections between 54 races and populations under different α.

	α=0 (177 predicted connections)				α=0.5 (161 predicted connections)				α=1 (144 predicted connections)			
	Node	Node	Simil. Win	Simil.	Node	Node	Simil. Win	Simil.	Node	Node	Simil.	Simil.
The first ten predictions	36	47	0.2046	0.5261	19	21	0.1888	0.5132	19	21	0.2352	0.5132
	36	45	0.1969	0.5253	12	21	0.1655	0.4229	12	21	0.1963	0.4229
	23	39	0.1938	0.5064	36	47	0.1608	0.5261	18	21	0.1854	0.3973
	33	47	0.1879	0.5082	33	47	0.1542	0.5082	19	20	0.1766	0.3854
	4	8	0.1825	0.4208	23	39	0.1502	0.5064	18	20	0.1613	0.3457
	23	32	0.1682	0.4068	36	45	0.1485	0.5253	7	29	0.1601	0.4984
	36	53	0.1584	0.5198	19	20	0.1458	0.3854	5	21	0.15	0.3177
	9	10	0.1576	0.4977	4	8	0.1456	0.4208	11	21	0.1471	0.3115
	1	8	0.149	0.4824	18	21	0.1417	0.3973	5	20	0.1458	0.3088
	10	21	0.1439	0.2564	14	21	0.139	0.3069	12	20	0.1445	0.3112
The last ten predictions	17	26	0.0085	0.1317	3	21	0.0089	0.0669	4	17	0.0086	0.4443
	18	46	0.0051	0.3086	18	47	0.0075	0.3548	12	26	0.0078	0.2327
	15	45	0.0037	0.3155	1	31	0.0072	0.3487	15	48	0.0069	0.3774
	10	29	0.0035	0.3364	7	20	0.007	0.0046	3	6	0.0059	0.0165
	31	32	0.0032	0.4706	10	24	0.0065	0.2771	25	30	0.004	0.3044
	18	54	0.0031	-0.0094	19	24	0.0063	0.2838	37	54	0.0037	0.0746
	1	17	0.0013	0.3344	4	18	0.0058	0.3573	31	54	0.0031	0.069
	11	32	0.0008	0.1182	3	7	0.0043	0.2998	18	35	0.0019	0.387
	6	24	0.0004	0.0246	15	37	0.0034	0.2005	32	42	0.0015	0.4077
	43	54	0.0003	-0.0906	18	37	0.0018	0.2065	4	29	0.0002	0.1191

Node IDs from 1 to 54 represent Lahu-China, Dai-China, Yao-China, Guangdong Han-China, Dulong-China, Buyi-China, Thais, Yi-China, Hunan Han-China, Southern Han-China, Singapore Han-Singapore, Pumi-China, Shanghai Han-China, Liaoning Han-China, Shegyang Han-China, Northwest Han-China, Northern Han-China, Manchu-China, Japanese, Hokkaido-Japan, Uighur-China, Kazak-China, Siberian Nivkhs population, Siberian Udegeys population, Siberian Koryaks population, Siberian Eskimo, Siberian Chukchi population, South American Indians Ticuna, South American Indians Terena, Siberian Evenki population, Siberian Kets population, USA whites, Spanish, German, Romanians, Bulgarian, Greek, Polish, Turks, Macedonians, Israeli Arabs, Iranian Jews, Ashkenazi Jews-Germany, Libyan Jews, Moroccan Jews, Ethiopian Jews, Native population-Australia's central desert, Yuendumu Native population-Australia, Kimberley native population-Australia, Cape York native population-Australia, North American blacks, and South African blacks.

3 Application Example

3.1 Analysis of 54 human races and populations and 14 common HLA-DRB1 alleles

Data of the world’s 54 human races and populations (nodes) and 14 common HLA-DRB1 alleles (attributes) (54×14; HLA_DRB1.txt; supplementary material) are from Jia (2001) (Zhang and Qi, 2014). In addition, an adjacency matrix (54×54; HLA_DRB1_adj.txt; supplementary material) and its two array form (×2, HLA_DRB1_adj_twoarrayform.txt; supplementary material) for the network of 54 human races (nodes), derived from linear correlation analysis, is given. In present example, I use Pearson correlation measure. Some results are given in Table 1.

Table 2 Predictions of connections between 12 Chinese populations under different α.

α=0 (25 predicted connections)				α=0.5 (24 predicted connections)				α=1 (23 predicted connections)			
Node	Node	Simil. Win	Simil.	Node	Node	Simil. Win	Simil.	Node	Node	Simil. Win	Simil.
3	9	0.1952	0.2517	**7**	**9**	**0.2172**	**0.5798**	**7**	**9**	**0.2577**	**0.5798**
4	9	0.1878	0.3699	**6**	**12**	**0.2072**	**0.58**	**6**	**12**	**0.2379**	**0.58**
6	9	0.184	0.4077	5	9	0.1816	0.41	**5**	**12**	**0.1795**	**0.4577**
5	9	0.1839	0.41	6	9	0.1812	0.4077	5	9	0.1794	0.41
5	**12**	**0.1794**	**0.4577**	**5**	**12**	**0.1795**	**0.4577**	6	9	0.1784	0.4077
7	**9**	**0.1766**	**0.5798**	4	9	0.1748	0.3699	10	12	0.173	0.4036
6	**12**	**0.1765**	**0.58**	10	12	0.1629	0.4036	4	9	0.1619	0.3699
2	9	0.1642	0.345	**10**	**11**	**0.1536**	**0.4698**	**10**	**11**	**0.1566**	**0.4698**
10	12	0.1528	0.4036	3	9	0.1527	0.2517	4	12	0.1461	0.3966
10	**11**	**0.1506**	**0.4698**	7	12	0.1326	0.2787	7	12	0.1239	0.2787
8	11	0.1467	0.32	2	9	0.1318	0.345	7	11	0.1198	0.3295
7	12	0.1413	0.2787	4	12	0.1318	0.3966	8	11	0.1164	0.32
8	12	0.1384	0.183	8	11	0.1315	0.32	**4**	**11**	**0.1162**	**0.4832**
7	11	0.1367	0.3295	7	11	0.1283	0.3295	3	9	0.1101	0.2517
4	**11**	**0.1302**	**0.4832**	**4**	**11**	**0.1232**	**0.4832**	2	9	0.0995	0.345
4	12	0.1176	0.3966	8	12	0.1099	0.183	8	12	0.0813	0.183
2	**10**	**0.1159**	**0.4808**	3	12	0.095	0.2429	3	12	0.0762	0.2429
3	12	0.1138	0.2429	3	11	0.0754	0.3054	3	11	0.0441	0.3054
3	11	0.1067	0.3054	**2**	**10**	**0.0686**	**0.4808**	9	11	0.0402	0.3214
1	9	0.1046	0.3837	1	9	0.059	0.3837	**2**	**6**	**0.036**	**0.5367**
2	**6**	**0.0695**	**0.5367**	**2**	**6**	**0.0528**	**0.5367**	9	12	0.0321	0.1284
2	5	0.0662	0.3984	9	11	0.0401	0.3214	**2**	**10**	**0.0214**	**0.4808**
9	12	0.0457	0.1284	9	12	0.0389	0.1284	1	9	0.0134	0.3837
9	11	0.04	0.3214	2	5	0.0166	0.3984				
2	12	0.0342	0.004								

Node IDs from 1 to 12 represent Tibetan, Uighur, Kazak, Xingjiang Han, Taiwanese, Hong Kong, Northern Han, Shanghai Han, Hunan Han, Manchu, Buyi, and Dai. Successfully predicted missing connections are in bold.

Using a finer adjacency matrix (54×54; HLA_DRB1_adj_Finer.txt; supplementary material) as the “complete network”, it is found that there are 46 missing connections in the incomplete network, HLA_DRB1_adj.txt. Using Pearson correlation, 32 (69.6%), 30 (65.2%), and 30 (65.2%) connections of missing connections for α=0, 0.5, 1 respectively, are successfully predicted. These connections are from 177, 161, and 144 predicted connections for α=0, 0.5, and 1 respectively. In addition, there are known 255 connections in the incomplete network, HLA_DRB1_adj.txt (potentially maximal 1431 connections). Thus only 12.4% (177/1431), 11.3% (161/1431), and 10.1% (144/1431) of possible connections are needed to be screened for further confirmation respectively, which greatly reduce the cost for experiments and observations.

3.2 Analysis of 12 Chinese human populations and 17 HLA-DQB1 alleles

Data of the 12 Chinese human populations (nodes) and 17 common HLA-DQB1 alleles (attributes) (12×17; HLA_DQB1.txt; supplementary material) are from Geng et al. (1995), Chang et al. (1997), Mizuki et al. (1997, 1998), et al. An adjacency matrix (12×12; HLA_DQB1_adj.txt; supplementary material) for the network of 12 human populations, derived from linear correlation analysis, is given. Use Pearson correlation measure and results are given in Table 2.

Using a finer adjacency matrix (12×12; HLA_DQB1_adj_Finer.txt; supplementary material) as the "complete network", it is found that there are 12 missing connections in the incomplete network, HLA_DQB1_adj.txt. Using Pearson correlation, 7 (58.3%), 7 (58.3%), and 7 (58.3%) connections of missing connections for α=0, 0.5, 1 respectively, are successfully predicted. In addition, these connections are from 25, 24, and 23 predicted connections for α=0, 0.5, and 1 respectively.

4 Discussion

It should be noted that the "complete networks" (with finer adjacency matrices) are defined in a relative sense. More complete networks may exist. Thus the percentage of successfully predicted missing connections may increase with further fining of networks.

In present study, we set $z^{ik}\geq 0$, to calculate $y^{ik}=z^{ik}/2$. However, the threshold can be lowered, for example, $z^{ik}\geq -h$, to calculate $y^{ik}=z^{ik}/2$, where $h>0$ is a constant. By doing this, the percentage of successfully predicted connections can be further increased (e.g., 70%, 80%, etc). But at the same time, the percentage of possible connections needed to be screened for further confirmation rises also, which will increase the cost for experiments and observations. Therefore, a compromise between the two percentages is unavoidable and a perfect prediction method is almost impossible.

The present algorithm is based on observed connections and the attributes of nodes. Two extreme situations can be reached by adjusting the α. α=0 (situation A) means only the similarities between a node (v_k) and the adjacent nodes of another node (v_i) being prepared to connect to v_k are considered; α=1 (situation B) means the comparison of the similarity between a node (v_k) and another node (v_i) being prepared to connect to v_k, and the similarities between v_i and its adjacent nodes, will be made. In practical applications, situation A sometimes occurs. If the mechanism for relationship between node-similarity and connection likelihood is unknown or unsure, α=0 is mostly suggested for use. In addition, (negative) Euclidean distance measure can be used in specific cases only, for example, node attributes are spatial coordinates.

The present algorithm is useful to not only the structurally stable networks but also evolving, structurally unstable networks. For structurally stable networks, Lü et al. (2015) proposed a prediction method, structural perturbation method, which was reported to be superior to the known hierarchical structure method (Clauset et al., 2008).

The effectiveness of the present algorithm depends on node attributes and similarity measures. Therefore, future works to improve the present algorithm should mainly focus on (1) selection of the key attributes of nodes in determining connection likelihood, and (2) addition of more specific similarity measures in the algorithm.

Making a little revision on the Matlab codes, the present algorithm can be used to predict which nodes might be connected by a new added node. Further, it can be used to describe network generation and evolution.

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, Discovery and Crucial Node Analysis of Important Biological and Social Networks (2015.6-2020.6), from Yangling Institute of Modern Agricultural Standardization, and Project on Undergraduate Teaching Reform (2015.7-2017.7), from Sun Yat-sen University, China.

References

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

Chang YW, Hawkins BR. 1997. HLA Class I and Class II frequencies of a Hong Kong Chinese population based on bone marrow donor registry data. Human Immunology, 56: 125-135

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

Geng L, Imanishi T, Tokunaga K, et al. 1995. Determination of HLA class II alleles by genotyping in a Manchu population in the northern part of China and its relationship with Han and Japanese populations. Tissue Antigens, 46: 111-116

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

Jia ZJ. 2001. Polymorphism of HLA-DRB1 gene in southern Chinese populations. PhD Thesis. 46-47, Sun Yat-sen University, Guangzhou, China

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

Mizuki N, Ohno S, Ando H et al. 1998. Major histocompatibility complex class II alleles in an Uygur population in the Silk Route of Northwest China. Tissue Antigens, 51: 287-292

Mizuki N, Ohno S, Sato T et al. 1997. Major histocompatibility complex class II alleles in Kazak and Han populations in the Silk Route of Northwest China. Tissue Antigens, 50: 527-534

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. Several mathematical methods for identifying crucial nodes in networks. Network Biology, 2(4): 121-126

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

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

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, Qi YH. 2014. Pattern classification of HLA-DRB1 alleles, human races and populations: Application of self-organizing competitive neural network. Selforganizology, 1(3-4): 138-142

Zhang WJ, Qi YH, Zhang ZG. 2014. Two-dimensional ordered cluster analysis of component groups in self-organization. Selforganizology, 1(2): 62-77

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