Article

A mathematical model for dynamics of occurrence probability of missing links in predicted missing link list

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 December 2015; *Accepted 15 February 2016*; *Published 1 December 2016*

Abstract
In most of the link prediction methods, all predicted missing links are ranked according to their scores. In the practical application of prediction results, starting from the first link that has the highest score in the ranking list, we verify each link one by one through experiments or other ways. Nevertheless, how to find an occurrence pattern of true missing links in the ranking list has seldomly reported. In present study, I proposed a mathematical model for relationship between cumulative number of predicted true missing links (y) and cumulative number of predicted missing links (x): $y=K(1-e^{-rx/K})$, where K is the expected total number of true missing links, and r is the intrinsic (maximum) occurrence probability of true missing links. It can be used to predict the changes of occurrence probability of true missing links, assess the effectiveness of a prediction method, and help find the mechanism of link missing in the network. The model was validated by six prediction methods using the data of tumor pathways.

Keywords mathematical model; missing links; prediction; occurrence probability; tumor pathways.

Network Pharmacology
ISSN 2415-1084
URL: http://www.iaees.org/publications/journals/np/online-version.asp
RSS: http://www.iaees.org/publications/journals/np/rss.xml
E-mail: networkpharmacology@iaees.org
Editor-in-Chief: WenJun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

Link prediction is conducted 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), which is expected to reduce the experimental costs for link finding. So far, numerous research on link prediction have been conducted (Clauset et al., 2008; Guimera and Sales-Pardo, 2009; Lü and Zhou, 2011; Lü et al., 2012; Barzel and Barabási, 2013; Bastiaens et al., 2015; Lü et al., 2015; Zhang, 2007, 2011, 2012a-c, 2013, 2015a-d, 2016a-d; Zhang and Li, 2015; Zhao et al., 2015; Zhou, 2015). Known so many prediction methods, each of them has its own mechanism for link generation, and their predictions for the same dataset are diverse. For example, the random perturbation method of Zhang (2016a) was developed based on degree growth of scale-free. Power-law generation method assumed that the degree of most networks is power-law distributed and power-law

distribution was used to fit degree distribution and predict msiing links (Zhang, 2016c). CN was based on common neighbors of two links (Lorrain and White, 1971). In most of the link prediction methods, all predicted missing links are ranked based on their scores. In the application of prediction results, starting from the first link, we usually verify each link through experiments or other ways. How to find an occurrence pattern of true missing links in the ranking list for reducing verification cost has seldomly reported. In present study, I tried to propose a mathematical model for relationship between cumulative number of predicted true missing links and cumulative number of predicted missing links in order to further enhance prediction efficiency and identify the mechanism of link missing.

2 Methods

2.1 Mathematical model

In the most of prediction methods for missing links, for a link being predicted, calculate the score (likelihood, probability, etc. A higher score means the greater probability of being a true missing link) of the link using the prediction method, and rank all predicted missing links according to their scores, from greater to smaller ones. The top links in the ranking list are more likely true missing links. The ranking list is always stored in a file with three columns, *A*, *B* and *C*. *A* stores IDs of "from" nodes, *B* stores IDs of "to" nodes, and *C* stores scores of predicted missing links.

First, define the occurrence probability of true missing links as

$$p=\mathrm{d}y/\mathrm{d}x \qquad (1)$$

where y is the cumulative number of predicted true missing links, and x is the cumulative number of predicted missing links, starting from the 1st predicted link with the highest score in the ranking list. A preceding link in the ranking list is more likely a true missing link than its succedent links. As a consequence, the occurrence probability of true missing links, p, will decline (from the intrinsic occurrence probability, i.e., maximum occurrence probability, r) to zero as the increase of cumulative number of predicted true missing links (y) until the expected total number of true missing links, K, is achieved. Therefore, as the first-order approximation of equation (1), let

$$p=r(K-y)/K \qquad (2)$$

and we have

$$\mathrm{d}y/\mathrm{d}x=r(K-y)/K \qquad (3)$$

where $K>0$, $0<r\leq 1$. Solving equation (3), the mathematical model for relationship between cumulative number of predicted true missing links (y) and cumulative number of predicted missing links (x) is achieved as

$$y=K(1-\mathrm{e}^{-rx/K}) \qquad (4)$$

According to the model (4), the cumulative number of predicted true missing links (y) increases as the increase of predicted missing links (x) in the ranking list, and tends to an asymptote, i.e., expected total number of true missing links, K (Fig. 1).

Model (4) can be used to predict the changes of occurrence probability of true missing links in the ranking

list. Based on the model (4), the expected total number of true missing links, *K*, and the intrinsic occurrence probability of true missing links, *r*, can be obtained by using data fitting.

The following are Matlab codes, linkPredModel.m, to obtain the relationship between *y* and *x* and the numerical method to obtain *K* and *r* by fitting relationship between *y* and *x*

```
scores=input('input the excel file name of scores: ','s');
misslinks=input('input the excel file name of missing links: ','s');
scores=xlsread(scores); misslinks=xlsread(misslinks);
n=size(scores,1); m=size(misslinks,1);
scores=sortrows(scores,-3);
x=1:n;
y=zeros(1,n);
ma=0;
for i=1:n
for j=1:m
if ((scores(i,1)==misslinks(j,1)) & (scores(i,2)==misslinks(j,2))) y(i)=ma+1; ma=y(i);break; end
if ((scores(i,1)==misslinks(j,2)) & (scores(i,2)==misslinks(j,1))) y(i)=ma+1; ma=y(i); break; end
end
y(i)=ma;
end
plot(x,y,'-');
xlabel('Cumulative number of predicted missing links (x)');
ylabel('Cumulative number of predicted true missing links (y)');
disp('Cumu. number of predi. missing links (x)          Cumu. number of predi. true missing links (y)')
[x'     y']
k=input('Input the estimated value of parameter K (e.g., 15): ');
r=input('Input the estimated value of parameter r (e.g., 0.1): ');
sig=input('Input the significanc level (e.g., 0.01): ');
beta=[k r];
[beta,R,J,SIGMA,MSE]=nlinfit(x,y,@predictfunction,beta);
K=beta(1)
r=beta(2)
deltabeta=nlparci(beta,R,J);
fitted=predictfunction(beta,x);
chi_square=sum((y-fitted).^2./fitted)
p=chi2cdf(chi_square,n-2)
if (p<sig) disp('The data fit model well at the given significanc level.');
else disp('The data is not able to fit model at the given significanc level.');
end
```

The following is the function predictFunction.m

```
function f=predictfunction(beta,x)
f=beta(1)*(1-exp(-beta(2)/beta(1)*x));
```

The software and data can be found in supplementary material of the present article.

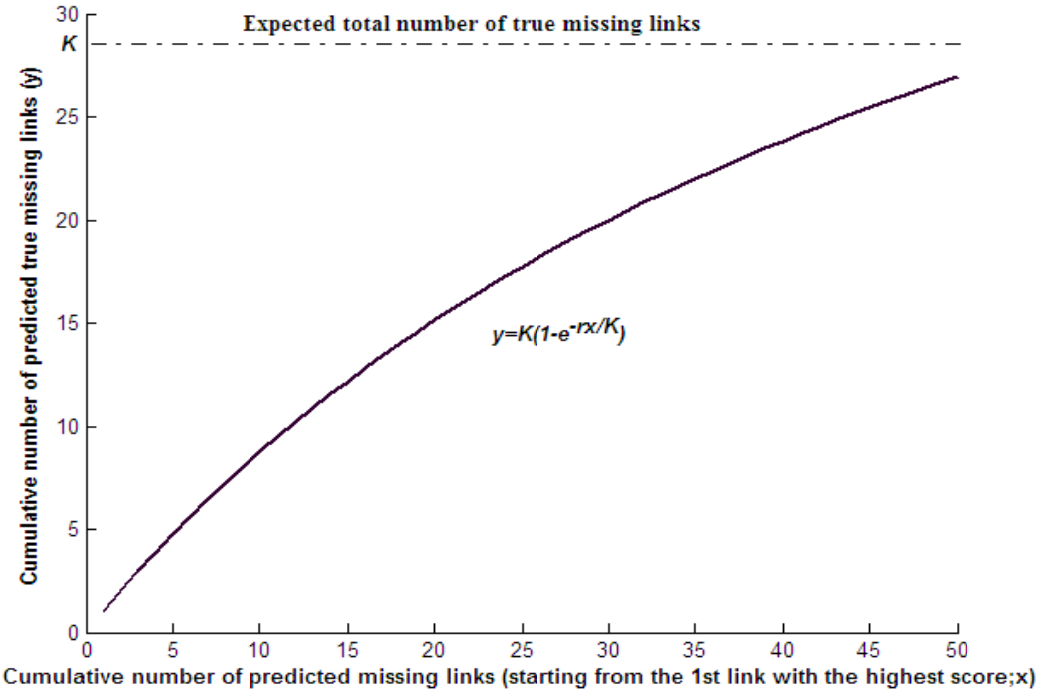

Fig. 1 Theoretical relationship between predicted true missing links and predicted missing links in the ranking list.

2.2 Implication and significance of model (4)

Suppose there are totally n predicted missing links in the ranking list. With the same total number (S) of predicted true missing links in the ranking list, the method with the larger value of r is a better method (correspondingly, with the same value of r, the method with the larger S is a better method). Suppose the total number of true missing links, which is unknown, is L. Obviously, K is the approximation of L, and $S \leq L$. As $r \to 1$, and $S \to L$, the prediction method tends to be a better method. Theoretically, r and K can be jointly used to assess performance of prediction methods. However, the estimated K may not well approximate L due to data quality for parameter fitting, and S is unknown at the early stage of link verification. Therefore, the parameter, r, is more reliable for use. In summary, main implication and application of model (4) include

(1) Intrinsic occurrence probability of true missing links (r) denotes the prediction capability or effectiveness of a prediction method and can thus be used to assess the performance of a prediction method.

(2) It is naturally concluded that the mechanism of link missing for the better method in the assessment is more likely the true mechanism of link missing.

(3) Model (4) (exactly, equation (2)) can be used to predict or estimate the occurrence probability of true missing links.

2.3 Prediction methods

In present study, six methods are used to predict missing links: (1) Random perturbation (Zhang, 2016a); (2) Power-law generation (Zhang, 2016c); (3) CN (common neighbors) (Lorrain and White, 1971); (4) Adamic-Adar (Adamic LA, Adar E. 2003); (5) RA (resource allocation; Zhou et al., 2009), and (6) Katz (Katz, 1953).

I use the data of original tumor pathways FAS, JAK-STAT, JNK, MARK and *p*53 (ABCAM, 2012; Huang and Zhang, 2012; Li and Zhang, 2013; Pathway Central, 2012; Zhang, 2016a; *pathwayname*.xls in

supplementary material). Links are removed from tumor pathways and the remaining (see missinglinks*pathwayname*.xls and *pathwayname*_Training.xls in supplementary material) is used to generate the training adjacency matrix.

Missing links are removed from original pathways following the inverse evolution process of the networks with power-law degree distribution. Therefore, the links missed following a known mechanism and the power-law based nethods (Zhang, 2016c), in particular, random perturbation (Zhang, 2016a) are the methods featured by this mechanism. Most networks have the degree distribution of power-law (Barabasi and Albert, 1999; Zhang, 2012a, 2016a, 2016c). For network evolution based link prediction, it is a reasonable and exact treatment.

The prediction results of missing links in tumor pathways, by six methods, can be found in separate directories of supplementary material (PrediScores_*methodname*.xls).

3 Results

3.1 Validation of the mathematical model

Compared with Fig. 1 and Fig. 2, a general coincidence between theoretical curve and observed curves can be founded. As indicated in Table 1, parameter fitting and statistic test further validate the mathematical model (4) that describes the relationship between cumulative number of predicted true missing links (y) and cumulative number of predicted missing links (x). In addition, the results show that the parameter r is more reliable than K.

Table 1 Model fitting and statistic test of missing link predictions of six methods for five tumor pathways.

Pathways	Model para./Chi square	Random perturbation	Power-law generation	CN	AA	RA	Katz
FAS (L=12 true missing links	K	9.2480	28.2614	-2.6×10^{9}	1.2×10^{9}	1.2×10^{9}	4.0375
	r	**0.0458**	0.0177	0.0141	0.0142	0.0142	0.0190
	χ^2	330.5661^{**}	92.2104^{**}	57.1607^{**}	169.0797	169.0797	61.7429^{**}
JAK-STAT (L=11 true missing links	K	7.2210	13.9050	3.8933	4.1197	4.1197	5.2×10^{6}
	r	0.0158	0.0250	**0.2368**	0.2042	0.2042	0.0070
	χ^2	36.1173^{**}	54.3138^{**}	5.3417^{**}	20.6666^{**}	20.6666^{**}	139.9072^{**}
JNK (L=16 true missing links	K	16.0194	21.8450	2.2015	4.4009	5.5020	6.8×10^{5}
	r	**0.1672**	0.0347	0.0359	0.0147	0.0138	0.0172
	χ^2	43.4027^{**}	127.4522^{**}	12.1910^{**}	21.7042^{**}	21.3651^{**}	695.3245^{*}
MARK (L=15 true missing links	K	3.2×10^{5}	4.9×10^{6}	2.2247	2.1745	2.1745	15.7222
	r	0.0104	0.0119	**0.0460**	0.0296	0.0296	0.0122
	χ^2	117.8132^{**}	71.5835^{**}	36.0525^{**}	69.4835^{**}	69.4835^{**}	496.0534^{**}
p53 (L=13 true missing links	K	11.2094	15.0853	1.7×10^{10}	3.0554	3.0530	6.8874
	r	0.0788	0.0163	0.0039	0.2080	**0.2250**	0.0327
	χ^2	59.8453^{**}	653.3348^{**}	56.2755^{**}	6.8479^{**}	7.4530^{**}	26.0670^{**}

**: p<0.01; *: p<0.05.

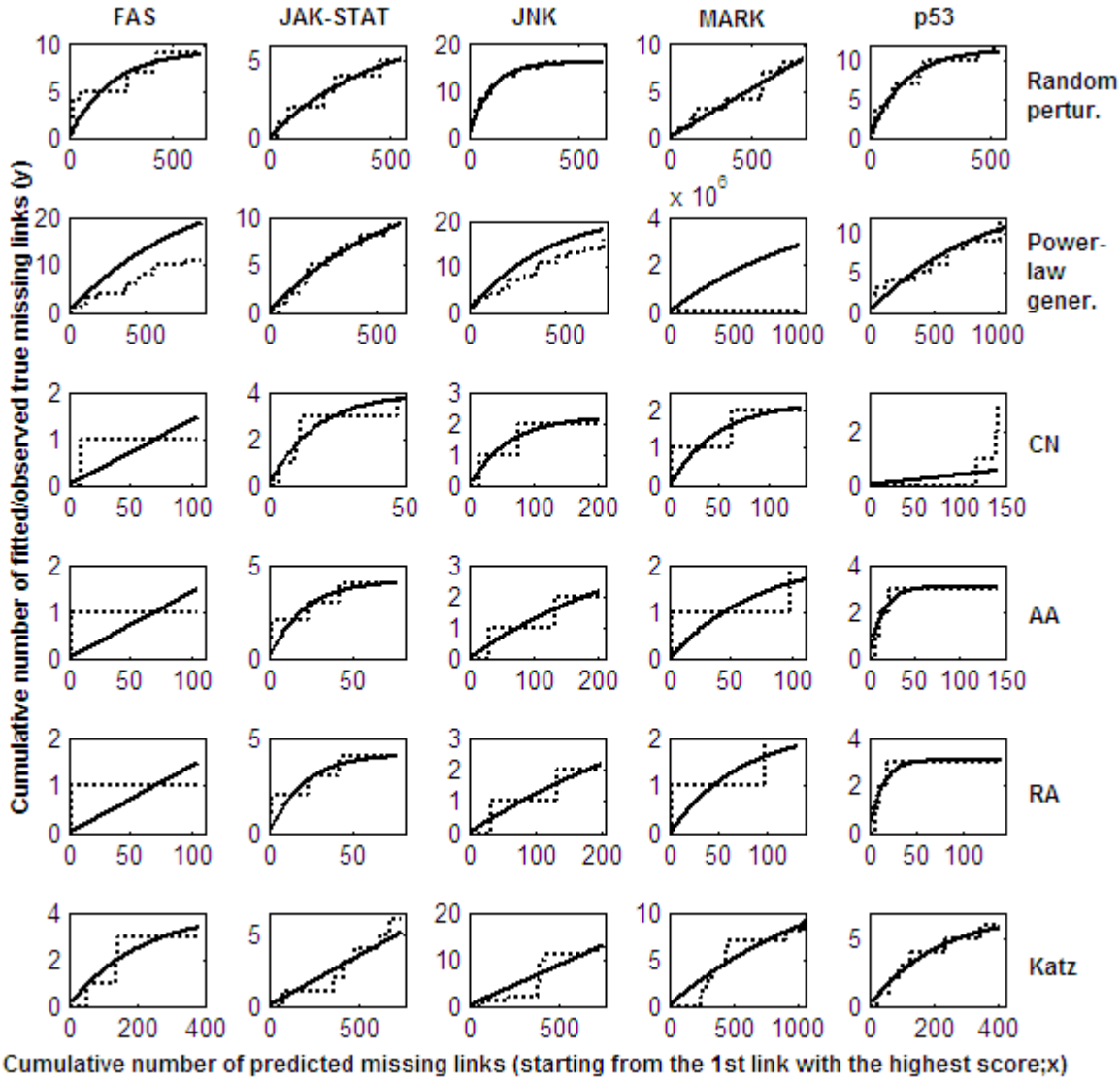

Fig. 2 Fitted and observed relationships based on model (4) (●●●: observed; ▬: fitted).

Table 2 Summary of prediction results of six methods for missing links in five tumor pathways.

Pathways	Type of links	Random perturbation	Power-law generation	CN	AA	RA	Katz
FAS	Predicted missing links (*n*)	**640**	866	105	105	105	386
	Predicted true missing links (*S*)	**9**	11	1	1	1	3
JAK-STAT	Predicted missing links (*n*)	545	613	**76**	**76**	**76**	736
	Predicted true missing links (*S*)	5	9	**4**	**4**	**4**	6
JNK	Predicted missing links (*n*)	**612**	709	201	201	201	759
	Predicted true missing links (*S*)	**16**	16	2	2	2	12
MARK	Predicted missing links (*n*)	819	1006	133	133	133	1166
	Predicted true missing links (*S*)	8	11	2	2	2	9
*p*53	Predicted missing links (*n*)	**539**	1067	142	142	142	406
	Predicted true missing link (*S*)	**12**	13	3	3	3	6

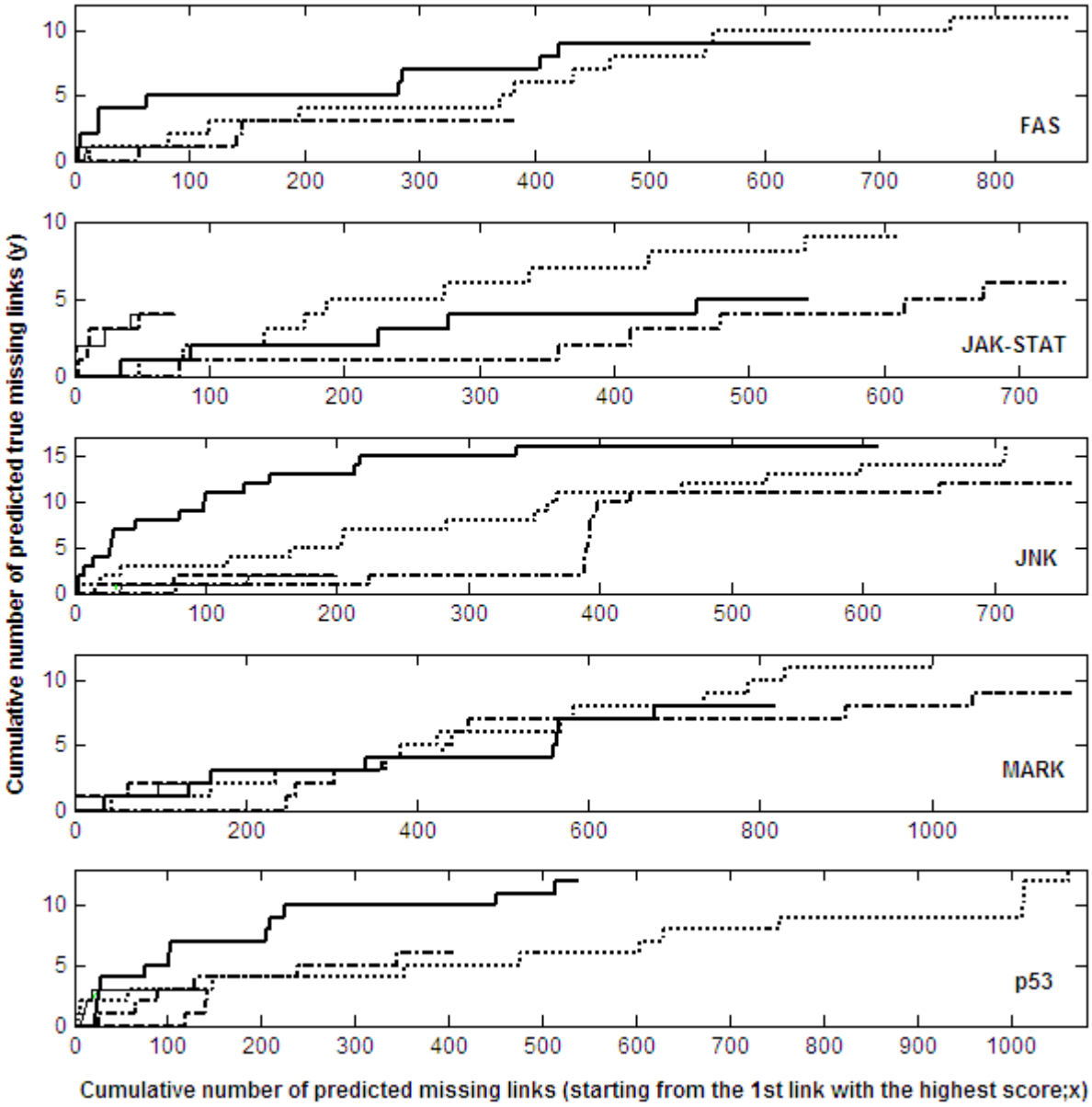

Fig. 3 Comparisons of relationships between six methods under five tumor pathways (●●●: Power-law generation; ——: Random perturbation; —●—: Katz; — — —: CN; ——: RA; —·—: AA). For FAS and JNK prediction, Random perturbation is significantly better than other methods, with the larger *r* values (steeper and earlier sprouted curves) and overall the larger total number of predicted true missing links than other methods, in exception of Power-law generation.

3.2 Judgement of prediction methods and discovery of link missing mechanism

A comprehensive survey on Table 1, Table 2 and Fig. 2, in terms of *r* and total numer of true missing links *S*, proves that Random perturbation performs significantly better than other methods, seconded by CN. The curves of Random perturbation (Fig. 2) are more similar to the theoretical curve in Fig. 1. The mechanism of link missing in Random perturbation is forseeablly the true mechanism.

4 Discussion

I guess that the pattern described by model (4) may has resulted from power-law distribution of true missing links in the ranking list, i.e., the occurrence probability of true missing links in the ranking list follows the power-law distribution.

Most of prediction methods were based on static topological structure only. Network evolution based (Zhang, 2012a, 2012c, 2015a, 2016b, 2016e), node similarity based (Zhang, 2015d), and sampling based (correlation based; Zhang, 2007, 2011, 2012b, 2013, 2015b; Zhang and Li, 2015) methods are used also. Model (4) in present study is suitable to predictions of all these types of methods.

To simply obtain K, we may carefully choose three points, (x_1, y_1), (x_2, y_2), (x_3, y_3), where $x_3-x_2=x_2-x_1$. Derived from equation (3), we have the estimate of K as the following

$$K=(y_2\times y_2-y_1\times y_3)/(2y_2-y_1-y_3)$$

However, a sound method to obtain K and r from data fitting is urgently needed.

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, China.

References

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

Adamic LA, Adar E. 2003. Friends and neighbors on the web. Social Networks, 25(3): 211-230

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

Katz L. 1953. A new status index derived from sociometric index. Psychometrika, 1(18): 39-43

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

Lorrain F, White HC. 1971. Structural equivalence of individuals in social networks. Annual Review of Sociology, 1(1): 49-80

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

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 (Zhang WJ, 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 node degree dependent random perturbation method for prediction of missing links in the network. Network Biology, 6(1): 1-11

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

Zhang WJ. 2016c. Generate networks with power-law and exponential-law distributed degrees: with applications in link prediction of tumor pathways. Network Pharmacology, 1(1): 15-35

Zhang WJ. 2016d. Network pharmacology: A further description. Network Pharmacology, 1(1): 1-14

Zhang WJ. 2016e. 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

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)

Zhou T, Lu LY, Zhang YC. 2009. Predicting missing links via local information. European Physical Journal B, 71(4): 623-630