http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (9): 1811-1817, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.1811.1817

Sampling Network Motif Detection Algorithm Based on Subgraph Extending and Subgraph Support Value

Jiawei Luo and Fengrong Zhu
School of Information Science and Engineering, Hunan University,
No. 252, Lushan South Road, Changsha, 410082, China

Abstract: Network motifs play an important role in biological networks but the detection is computing complex and time consuming. Sampling method has been used in network motif detection to decrease calculated amount, however the inevitable sampling error influences the result validity seriously. In order to reduce the sampling error, a sub graph extending method is introduced to improve the computation performance and a sub graph support value is proposed to get more potential topology information of the network and the sub graph support value as a parameter is used to calculate the sub graph concentration of network. The experiment results indicated that the using of sub graph support value reduced the sampling error and this study achieved better computing performance and sampling stability.

Key words: Network motif, sampling algorithm, subgraph extending, subgraph support value

INTRODUCTION

One type of small connected subgraph which has significantly higher frequency in a network than in random networks is defined as network motif, it was proposed firstly in 2002 by Milo *et al.* (2002). Network motif is helpful to understand various complex networks (Kong and He, 2010) and is useful for the research of structure and function of biological network. The network motif study on transcriptional regualtion network revealed that the network motif have information processing feature (Shen-Orr *et al.*, 2002). And the application of network motifs in the prediction of interaction and function module finding shows good effects (Albert and Albert, 2004; Saito *et al.*, 2002a, b).

Although network motif detection is a very complex problem, the importance of it in bioinformatics urges researchers to take part in further studies (Qin *et al.*, 2009). The detection of network motif typically consist of three subtasks: the generation of random networks which have the same vertices degree sequence with the input network; subgraph isomorphism computation and classification; calculation of statistical metric and determining network motif (Wong *et al.*, 2011).

As the definition of the network motif described, subgraph frequencies are computed in both the real network and random networks. Generating random networks is one important part in detection of network motif. The familiar algorithm generates the random graphs by randomly switching edges between vertices from the original graph. This switching technique is never certain when proper randomization has been reached (Wong *et al.*, 2011), however from the statistical principle point of view, the generated networks are required to satisfy to the randomness enough.

Reducing the time consuming of the isomorphism testing is crucial to an efficiency network motif detection algorithm. Graph isomorphism is known as a NP-complete problem, the exponentially rise of computation time make it hard to deal with the big size graph (Foggia *et al.*, 2001). The best runtime of the n_node graphs of the known algorithm is $2^{O(\sqrt{n \log n})}$ (Babai and Codenotti, 2008; Johnson, 2005). Canonical labeling of the network's nodes is been used to solve the isomorphism. For example, one of canonical labeling-based isomorphism testing algorithms named NAUTY has been used in many network motif detection tools (Wong *et al.*, 2011). The GraphGen (Li *et al.*, 2007) algorithm divides the mining frequency subgraph into two parts of finding frequency of subtree and extending the subtree to the subgraph, it only need to compute the subtree isomorphism, also improved the performance of the algorithm.

In order to determining statistical significance of the frequency of a subgraph, it is necessary to get the appearance proportion of all types of n_node subgraphs in real network and random networks (Kashtan *et al.*,

Corresponding Author: Jiawei Luo, School of Information Science and Engineering, Hunan University, No. 252, Lushan South Road, Changsha, 410082, China Tel: 86-731-88821971 Fax: 86-731-88822417

2004). For the reason of computing time of enumerating methods would increase sharply as the graph size, Kashtan *et al.* (2004) proposed an edge sampling algorithm ESA. This method samples a set of subgraphs to estimate their frequency in network, the runtime of this method is not as the exhaustive method closely related to the graph size. Compare with the enumerate algorithm, ESA has great advantage of time consuming, however the edge sampling strategy leads to sampling bias that the possibility of each subgraph to be sampled is not equally (Wernicke, 2005). To correct the bias of sampling algorithm, A node sampling algorithm Rand_Esu (Wernicke, 2006) was presented and a tool named FANMOD is implemented based on it (Wernicke and Rasche, 2006). Rand_Esu used a node extension pattern growth tree makes all of the leaf subgraphs have the same possibility to be sampled. And the size of motifs Rand_Esu detected reached to eight vertices.

Sampling is a statistical method which deduces the ensemble distribution from sample indicators. There are two types of factors leads to the incorrect sample estimate. The one is nonsampling error, the reason of it is violated the random sampling principle (Jin *et al.*, 2002). Edge sampling strategy ESA used leads to oversampling of some subgraphs is belongs to the nonsampling error, however, it is possible to prevent. For instance, by using a node extension pattern growth tree, Rand_Esu eliminates the bias caused by edge sampling. The other one is sampling error which is inevitably even following the random sampling principle but it is controllable (Jin *et al.*, 2002). The evaluation of estimating of overall distribution is according to the sampling error, if the sampling error is large that means the estimating is incorrect (Jin *et al.*, 2002). Thus it is important to reduce the sampling error.

For the reason of inevitable sampling error and different subgraph distribution of various networks, sampling method leads to the estimating of the subgraph frequency apart from the original frequency, especially when sampling size is small. A method using subgraph extending and Subgraph Support Value (SSV) called SE&SSV was presented to reduce the sampling error, the subgraph extending can reduce the time consuming and the SSV can get more potential topological information of network to correct the error of sampling distribution, the experiments confirmed it.

MATERIALS AND METHODS

Networks

Real networks: The real network used in this study is an immunoglobulin protein network which contains 95 nodes, 213 edges, based on the PDB database (www.rcsb.org/pdb/) and it's PDB ID is 1A4J.

Random networks: The subgraphs sampling from the real networks have to compare with random networks to judge whether it is significance in the number of appearance. The random networks are required to have the same node degree sequence to the real network, usually, exchange several edges' start node and end node to keep the generated rand network have the same node degree sequence to the real network (Wong *et al.*, 2011), however, the randomness of generated network is not enough especially in highly dense PPI networks. Furthermore the nodes with higher degree are key node usually and may be the node within motif in greater probability, the edge exchange strategy would keep more edges of high degree node same to the real network, it will make the subgraphs sampling unsatisfied to the statistical principle enough. In this study, only undirected networks are considered, Another generating method is adopted to generate a series of random networks with the same degree sequence to the real network and could completely meet the randomness, details is described as Fig. 1.

Method

Subgraphs enumerating and sampling: Esu and Rand_Esu (Wernicke, 2005, 2006) are used to traverse the network and sampling the subgraphs in this study. The isomorphic subgraphs will be categorized as a same type, a set of n_node subgraphs with corresponding number of isomorphic subgraphs can be got, either in real network or random networks generated by the method described earlier.

Subgraph support value: In the step of isomorphism judgement, usually, only when a subgraph is isomorphic

Algorithm: Getting Random Network(G)(Get_RN)
Input: A real network G;
Output: A random network Gr;
1: Create a matrix Gr with the dimension same to G;
2: Get the node degree sequence(D) of G;
//D = {D_i| D_i is the degree of node i, i=0 to n-1;} n is the node number of real network;
3: While(exist nonzero element in D){
4: Select the maximum element D_m in D;
5: m = D_m;
6: For(j=0 to m){
7: Get a random integer k between [0, n) which $D_{k>0}$ and Gr[m, k] = 0;
8: Set Gr [m, k] = 1, Gr [k, m] = 1;
9: D_m--; D_k--;}
10: }
11: End;

Fig. 1: Algorithm of get random networks

to the other one, the isomorphic number of this type will plus one like ESA, Esu, Rand_Esu (Kashtan *et al*., 2004) (Wernicke, 2005, 2006). A Subgraph Support Value (SSV) is used in this study, subgraph support value is a probability to measure the isomorphic of two subgraphs emerge in network.

The computation of SSV was divided into two parts: The first part is when subgragh's node number is less than the farther graph's: Because of the (n-1)_node tree extend to n_node tree will add only one edge, so for simplicity using the subtree replace the subgraph when the node number is less than the farther graph. There is only one type of 3_node tree, so starting at 3_node tree. The support value of this part is called as Subtree Support Value (STV). The process of get STV from 3_node to 5_node tree is shown in Fig. 2.

There is only one type of 3_node tree, extend t^3 by add one node and edge can get two types of 4_node trees t_1^4 and t_2^4, we assume that all nodes have same probability to connect to the added node, so the probability of t_1^4 extend from t^3 is 1/3, t_2^4 is 2/3, then we can calculate the isomorphic probability of two 4_node trees extend from t^3:

$$STV_{3\_4} = (\frac{1}{3}\times\frac{1}{3})+(\frac{2}{3}\times\frac{2}{3}) = \frac{5}{9}$$

If extend from 4_node trees to 5_node trees, the 4_node tree have the probability of 1/3 to be t_1^4 and the probability of 2/3 to be t_2^4. Then the isomorphic probability of 5_node tree extend from 4_node tree:

$$STV_{4\_5} = \frac{2}{3}\times\left[\left(\frac{2}{4}\times\frac{2}{4}\right)+\left(\frac{2}{4}\times\frac{2}{4}\right)\right]+\frac{1}{3}\left[\left(\frac{1}{3}\times\frac{1}{3}\right)+\left(\frac{2}{3}\times\frac{2}{3}\right)\right] = \frac{14}{27}$$

can be calculated and:

$$STV_{5\_6} = \frac{2}{7}$$

$$STV_{6\_7} = \frac{57}{180}$$

alike and so on.

Before the second part, it is necessary to know the definition of inside edge and outside edge: in the process of graph extending, if the added edge introduces a node to the graph, the added edge was defined as outside edge, else defined as the inside edge (Li *et al*., 2007).

The second part is extending the generating tree to the graph which has the same edge number to the farther graph. When subtree is the generating tree of father graph, the extension is adding all inside edges. The number of different edge adding patterns without regard to the graph symmetry is used to calculate the probability

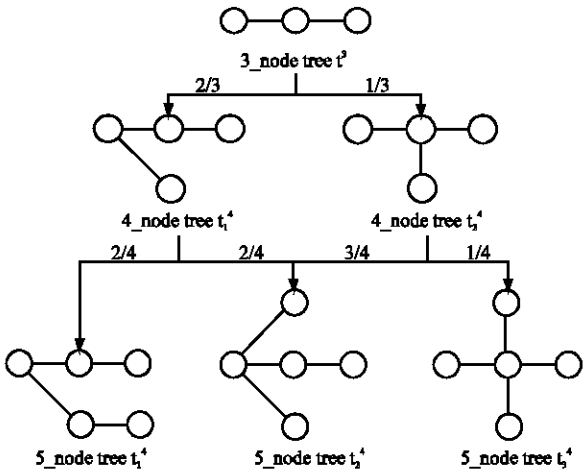

Fig. 2: Process of getting STV from Subtree extending

of different patterns. And the probability of different patterns is took as the support value of generating tree to father graph called generating tree support value (SGV). Formula (1) is given for the calculation of SGV:

$$SGV = \begin{cases} \left(\prod_{i=0}^{en-1}\frac{1}{c_n^2-ew-i}\right)^2 & \left(e = en+ew \le \frac{1}{2}\left[\frac{n(n-1)}{2}\right]\right) \\ \left(\prod_{i=0}^{\frac{n(n-1)}{2}-e}\frac{1}{c_n^2-ew-i}\right)^2 & \left(e = en+ew > \frac{1}{2}\left[\frac{n(n-1)}{2}\right]\right) \end{cases} \quad (1)$$

where, n is the node number of the subgraph, en is the inside edge number, ew is the outside edge number. Through two parts above SSV of two graphs can be got, the procedure is shown in Fig. 3.

As is an example given in Fig. 4 to describe how to get SSV, there are two unisomorphic 5_node subgraphs gi5 and gri5. The max size of isomorphic subtree they have is 4_node like t4 and tr4, For the reason of gi5 and gri5 is a proper subgraph of random network, at least there have another node connect to it. If we extend tr4 to 5_node tree tr5 by adding an external node and edge, the probability to get two isomorphic 5_node trees is STV_{t4_t5} = 5/9. And extend the 5_node tree by adding inside edges also have a certain probability to get two isomorphic graphs in this example are gi5 and gre5. In this situation there is only one inside edge, thus the probability to get two isomorphic graphs is:

$$SGV = \left(\frac{1}{C_3^2-4}\right)^2 = \frac{1}{36}$$

So the SSV of gi5 and gri5 is STV*SGV = 0.0154.

From the way of get SSV, it can be seen that the subgraph support value consider more potential

topological information of the network and could estimate the total distribution of subgraph in network more globality and accuracy than simply plus one, the experiments could confirm it. The other way round, it is possible to sample less subgraphs to estimating the distribution in order to improve the computation efficiency at the same accuracy level.

Subgraph extending: Isomorphic graphs at least have one type of isomorphic subgraph, for this reason a subgraph extending method is proposed to cut off the unisomorphic graphs by judging whether there is an isomorphic subgraph. And because of the subtree is the simplest subgraph and there is the least time consuming in computing isomorphic, the subtree isomorphism is took to prune unisomorphic graphs. Based on subgraph extending, subgraph support value is introduced to measure the significance of subgraphs. The description of subgraph extending and SSV algorithm is shown in Fig. 5.

Through subgraph extending a SSV is got to evaluate the isomorphic probability of two subgraphs. Then take the SSV to calculate subgraph concentration which described in next section. SSV considered the connection of subgraph's neighborhood include in network, it can get more potential topological information of the whole network.

```
Algorithm: Getting Subgraph Support Value(g, g')(Get_SSV)
1:  Input: Two graphs g and g', STV;
2:  Output: Subgraph support value SSV;
3:  Calculate the SGV of g and g' use the Formula (1);
4:  For(i = 4 to n-1){
5:  If (there is not isomorphic i_node subtree of g and g')
6:  Return
7:  }
8:  If (g isomorphic to g') Return 1;
9:  Else Return SGV;
10 End;
```

Fig. 3: Algorithm of get subgraph support value

Subgraph concentrations: The support value of appearance of subgraphs of type i is SPi. SPi is the total SSV of subgraph i got in network traverse. The concentration of n_node subgraphs of type i is the ratio between their support value and the total support value of n_node subgraphs in the network:

$$Ci = \frac{SP_i}{\sum_i SP_i} \quad (2)$$

The normalized value is used to measure the distribution proportion of one type of subgraph in random network, then compare with the proportion of it in real network to judge whether it is a network motif.

RESULTS

In order to evaluate the performance of our algorithm and Rand_Esu, both of the algorithms were implemented in Java and testified the superior performance from two aspects, the one is time consuming, the other is sampling accuracy and stability. All these tests were done on a computer with an AMD Athlon(tm) 7850 Dual_core Processor 2.80 GHz and with 2 Gb of memory.

The comparison of time consuming: Firstly on the time consuming, ran the program of subgraph extending without consider the SSV on the real network which was mentioned earlier. For the reason of edge number of subgraphs would influence the time consuming of isomorphic judging and the number of edge of each

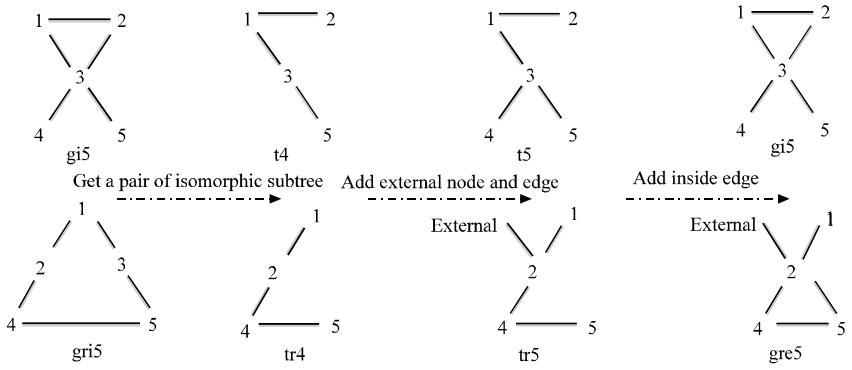

Fig. 4: An example of getting subgraph support value

Algorithm: Subgraph extending and SSV(G, k)(SE&SSV)
Intput: A network G, subgraph size k;
Output: A set of sampled subgraphs S = {SG_1, SG_2, …, SG_n} and corresponding SV = {SP_1, SP_2, …, SP_n};
1: Call Rand_Esu(G, k);
2: For each subgraph g get by Rand_Esu do{
3: For(i = 0 to size of S){
4: Call Get_SSV(g, SG_i) to get SSV_{g_SGi};
5: $Sp_1 = SP_i + SSV_{g\_SGi}$;}
6: If(there is not $SSV_{g\_SGi} = 1$)
7: Add the g into S as a new type and add the corresponding SP_g =
8: 1 into SV;
9: }
10: End;

Fig. 5: Algorithm of Subgraph Extending and SSV

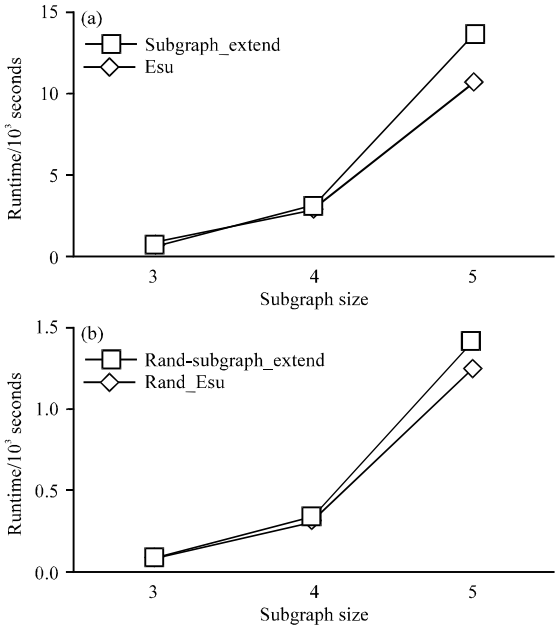

Fig. 6 (a-b): Comparison of (a) Enumerating runtime and (b) Sampling runtime

sampled graph is indeterminate, in order to reduce the impact of this, the average run time of ten times was took as the value to measure the algorithm performance, subgraphs in the size of 3, 4 and 5 have been executed in both of enumerating and sampling programs, the average run time is shown in Fig. 6.

Compare to the Esu, subgraph extending algorithm have excellent performance in run time either in enumerate or sampling method. Because there is only one type of 3_node tree subgraph extending algorithm degrade to Esu and when the size of subgraph is 4, because there is only two types of 4_node trees and 6 types of 4_node graphs, so there are few unisomorphic graphs have been cut off at the extending procedure, the improvement is not

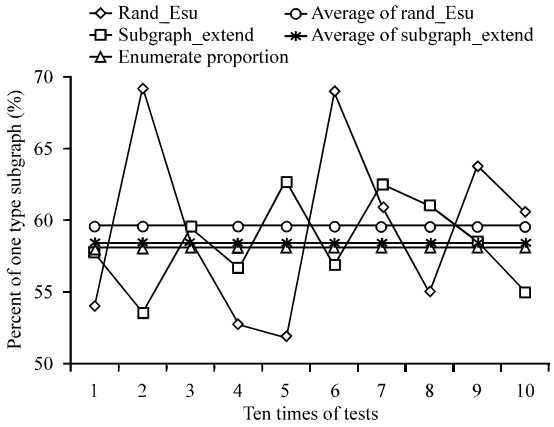

Fig. 7: Detail concentration distribution of 4_node subgraph tests

obviously. As the subgraph size increasing there are more unisomorphic graphs been cut off, accordingly the computing time reduced much more like the size 5 in the experiment. From Fig. 6a, the average runtime of Esu is about 13.6×10^3 seconds when the subgraph size is 5 and the average runtime of Subgraph_Extend is about 10.7×10^3 seconds that is 21% less than Esu's. Figure 6b shows the average runtime of sampling methods. The sampling probability was set as 0.1 in both methods. When the subgraph size is 5, the average runtime of Rand_Esu is about 1.4×10^3 seconds and of Rand_Subgraph_Extend is about 1.24×10^3 seconds, the average runtime of Rand_Subgraph_Extend is nearly 11.5% less than Rand_Esu's.

Performance on sampling accuracy and stability: Secondly compare the stability and accuracy. There are three groups of experiments were took to testify extending method more stable and accuracy, the type of subgraph with the highest concentration in real network was selected as the metric and record the concentration of this type of subgraph got by sampling algorithm each time. The concentration was calculated by Formula (2). Both of the methods were ran ten times at the subgraph size of 4, 5, 6. The detail parameters and results are shown in Table 1 and Fig. 7-9.

In Table 1, the column of enumerate proportion recorded the actually proportion of network that got by enumerate method. The column of average concentration recorded the average concentration of ten times tests of both Rand_Esu and Subgraph_Extend, the values closer to enumerate proportion means the corresponding method has better accuracy. The records of standard deviation column are used to evaluate sampling stability, the smaller value indicates the more stable performance.

Table 1: Parameters of experiments and the results of enumerate proportion, average concentration and standard deviation

Subgraph size	Total subgraphs	Sample subgraphs	Sample probability	Enumerate proportion (%)	Average concentration (%)	Standard deviation	Algorithm
4_node	2043	200	0.1	58.05	59.52	0.060	Rand_Esu
					58.35	0.029	Subgraph_Extend
5_node	6825	700	0.1	34.96	40.94	0.058	Rand_Esu
					35.64	0.031	Subgraph_Extend
6_node	23511	2000	0.1	20.05	27.66	0.046	Rand_Esu
					23.06	0.029	Subgraph_Extend

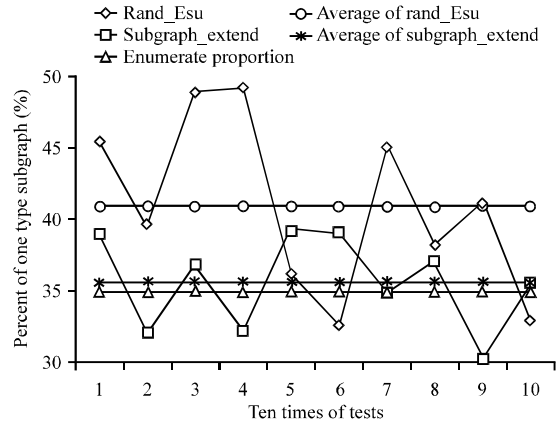

Fig. 8: Detail concentration distribution of 5_node subgraph tests

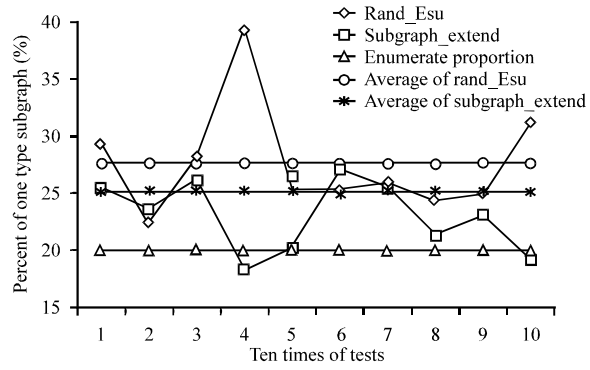

Fig. 9: Detail concentration distribution of 6_node subgraph tests

From Table 1, it can be seen that the average concentrations of subgraphs got from this study are closer to the actually proportion than Rand_Esu. And the standard deviations of our method are smaller, they are about a half of Rand_Esu's at 4_node and 5_node subgraph size and about two thirds of Rand_Esu's at 6_node subgraph size. It also can be seen from the Fig. 7-9 that the fluctuation of fold lines of our method is smaller than Rand_Esu's. It is proved that the SSV can get more potential topological information of whole network, this study can improve the accuracy of sampling algorithm and have more stable performance.

DISCUSSION

Sampling error is a factor that cannot be ignored in the sampling method whcih directly affect the correction of results. The common method to reduce the sampling error is expands the sample quantity, however it will relatively increase the amount of calculation (Jin *et al*., 2002). Kashtan *et al*. (2004) discussed the error ratio of ESA. The method they utilized to keep the error in a reasonable range is sample quantity expansion. The same to the Wernicke in discussion of Rand_Esu (Wernicke, 2005). To enhance the accuracy by expanding sample population is feasibility in some extent but the excessive expansion will lose the runtime superiority of sampling method. In this study, instead of enlarging the sampling population, a value of SSV that include more topological information is utilized to reduce the sampling error. And because of the subgraph extending and the easy calculation of SSV, reduction of sampling error is effective.

CONCLUSION

Because of the complex computation of network motif detection, subgraph sampling algorithm has been proposed. Sampling network motif detection is computing efficiency and could find larger motifs, however the accuracy of motifs it found is based on reasonable sampling error range. For the reason of uncertain of subgraph distribution of networks and simple random sampling will take in sampling error in large, a method of subgraph extending and introduce SSV in the computation of concentration has been proposed whcih can depress the sampling error. From the results of experiments it can be seen that our algorithm achieved the improved performance both in time consuming and sampling stability.

In future work, we will make further research to introduce more useful network characteristics and protein function information to improve the estimating accuracy of subgraph distribution, like the degree sequence of the nodes, essential proteins etc.

REFERENCES

Albert, I. and R. Albert, 2004. Conserved network motifs allow protein-protein interaction prediction. Bioinformatics, 20: 3346-3352.

Babai, L. and P. Codenotti, 2008. Isomorphism of hypergraphs of low rank in moderately exponential time. Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science, October 25-28, 2008, University of Chicago, USA., pp: 667-676.

Foggia, P., C. Sansone and M. Vento, 2001. A performance comparison of five algorithms for graph isomorphism. Proceedings of the 3rd IAPR-TC15 Workshop Graph-Based Representations in Pattern Recognition, May, 2001, Italy, pp: 188-199.

Jin, Y.J., Y. Jiang and X.Y. Li, 2002. Sampling Technique. Renmin University Press, China.

Johnson, D.S., 2005. The NP-completeness column. ACM Trans. Algorithms, 1: 160-176.

Kashtan, N., S. Itzkovitz, R. Milo and U. Alon, 2004. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs. Bioinformatics, 20: 1746-1758.

Kong, D. and J. He, 2010. Survey of motif discovery algorithms in protein-protein interaction networks. Comput. Technol. Dev., 20: 1-8.

Li, X.T., J.Z. Li and H. Gao, 2007. An efficient frequent subgraph mining algorithm. J. Software, 18: 2469-2480.

Milo, R., S. Shen-Orr, S. I tzkovitz, N. Kashtan, D. Chklovskii and U. Alon, 2002. Network motifs: Simple building blocks of complex networks. Science, 298: 824-827.

Qin, G., L. Gao and J. Hu, 2009. A review on algorithms for network motif discovery in biological networks. Acta Electronica Sinica, 37: 2258-2265.

Saito, R., H. Suzuki and Y. Hayashizaki, 2002a. Construction of reliable protein-protein interaction networks with a new interaction generality measure. Bioinformatics, 18: 756-763.

Saito, R., H. Suzuki and Y. Hayashizaki, 2002b. Interaction generality, a measurement to assess the reliability of a protein-protein interaction. Nucleic Acids Res., 30: 1163-1168.

Shen-Orr, S.S., R. Milo, S. Mangan and U. Alon, 2002. Network motifs in the transcriptional regulation network of *Escherichia coli*. Nature Genet., 31: 64-68.

Wernicke, S., 2005. A faster algorithm for detecting network motifs. Algorithms Bioinform., 3692: 165-177.

Wernicke, S. and F. Rasche, 2006. FANMOD: A tool for fast network motif detection. Bioinformatics, 22: 1152-1153.

Wernicke, S., 2006. Efficient detection of network motifs. IEEE/ACM Trans. Comput. Biol. Bioinf., 3: 347-359.

Wong, E., B. Baur, S. Quader and C.H. Huang, 2011. Biological network motif detection: Principles and practice. Briefings Bioinf., 13: 202-215.