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 (7): 1326-1333, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.1326.1333

A New Subgraph Sampling and Matching Algorithm for Probability Motifs Detection in PPI Networks

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

Abstract: Network motif detection is a very important problem in analysis of Protein-protein Interaction networks (PPI). In this study, an efficient algorithm for finding probability motifs in PPI networks is presented. First, a new sampling algorithm is provided to do subgraph mining; it is based on the adaptivity of the extension set of the subgraph. Then, both topological structure and biological significance are combined to do subgraph matching for calculating the mismatch point between subgraphs. Finally, the similar subgraphs are grouped by comparing to a mismatch threshold and the matrix of probability motif for each group will be calculated. Three PPI networks of Saccharomyces cerevisiae are used to test algorithm and achieve highly efficient and stable experimental results. Kinds of probability motifs are exactly found.

Key words: Network motif, sampling algorithm, subgraph matching, PPI networks

INTRODUCTION

For the past few years, with a large number of high throughput experiments and the development of using bioinformatics methods widely in the prediction of protein interactions field, people get more and more available Protein-protein Interaction network (PPI) data.

As a very important research field of bioinformatics, PPI network is used to discuss the evolution of biological system problem increasingly. Motifs are the building blocks of PPI networks and showing a very important local property of PPI networks. Milo *et al.* (2002) first provided this definition of PPI network motif. Motifs are the recurring and significant patterns of interconnections. These patterns have a much higher frequency of occurrences in real networks than in the random networks. It is so important, that the discovery and analysis of motifs caused the bioinformatics, complex network research and social statistics fields of wide concern. Network motifs in biological networks refer to some biological functions or do some information processing tasks (Alon, 2007), they can be used to predict protein-protein interaction (Albert and Albert, 2004) and discover underlying network decomposition (Itzkovitz *et al.*, 2005).

Later on, Berg and Lassig (2004) proposed a new definition: probability motif. They considered the motifs were not necessarily identical patterns and discussed motifs grouping of mutually similar subgraphs. They derived a scoring function to establish a statistical model for the occurrence of probability motifs and then they developed a search algorithm for matching motifs called graph alignment which was similar to sequence alignment (Xiang *et al.*, 2010). The algorithm designed by Berg and Lassig (2004) first introduced the concept of probability motif and this method didn't need to produce a large number of random networks just like traditional methods, so it had a higher efficiency. On the basis of Berg and Lassig (2004) and Jiang *et al.* (2006) thought the input network was also can be probability network. The algorithm further enlarged probability's scope, not only the motifs were not necessarily identical patterns but also the whole input network was uncertainly. All of the uncertainty theories are very accord with the real biological networks and gradually become the emphasis of bioinformatics research (Zou *et al.*, 2010).

Researchers developed kinds of algorithm for motif detection. ESA (edge sampling) algorithm was provided by Kashtan *et al.* (2004) ESA algorithm is independent from the network scale; the analysis of large network is effective to find bigger size motifs. But ESA algorithm cannot ensure getting all the subgraphs and the same subgraph may be multiple sampling. In order to prevent search repeated subgraphs, FANMOD: a tool for fast motif detection was proposed by Wernicke (2006) and Wernicke and Rasche (2006). A faster exhaustive

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

algorithm ESU (Enumerate Subgraph) was used by FANMOD to enumerate all size-k subgraphs and each subgraph appears only once. According to the larger experimental data, they also proposed a sampling algorithm Rand-ESU (Wernicke, 2006) that had been widely used in follow-up probability motifs research. NeMoFinder algorithm adopted the idea SPIN to search for repeated trees and extended to subgraphs, then counted the subgraph to ensure the motifs (Chen *et al*., 2006). NeMoFinder had enabled the discovery of network motifs with sizes ranging all the way to meso-scale but at the cost of missing some potentially interesting motifs (Ciriello and Guerra, 2008). How to determine the similarity of probability motifs is a key problem in this few years. Researchers matched the subgraphs with their adjacent matrix (Wong *et al*., 2011) or gave each subgraph a graph code based on the giving rule (Qin and Gao, 2012) to check their differences. The purpose was to find a suitable method for subgraph alignment then cluster the similar ones to find probability motifs.

In this study, a new sampling algorithm is firstly provided for subgraphs based on vertex adaptive rule. Then the study combines both topological structure and biological significance to do subgraphs matching. At last, from the results of front, a classification algorithm is used to cluster the similar subgraphs and then calculate the average of adjacency matrixes for each cluster to determine the probability motifs. Three PPI networks of Saccharomyces cerevisiae are used as the test data and achieve highly efficient and stable experimental results. Kinds of probability motifs are exactly found.

MATERIALS AND METHODS

Generally, finding probability motifs consists of three subtasks: (1) Find which subgraphs occur in the input graph and in which number of size-k, (2) Match the subgraphs with some rules to determine which of these subgraphs are similar, (3) Group the similar subgraphs into classes to calculate the probability motifs from PPI networks.

New subgraph sampling algorithm: As showing in the research of Wernicke (2006) the ESU-tree reflects the algorithm of all processes. The algorithm was applied to mark all the vertexes and sort them, then program began in a single vertex, add a vertex each iteration until the subgraph to the desired size-k. In addition to have set $V_{Subgraph}$ like ESA (Kashtan *et al*., 2004), there was also another set called $V_{Extension}$. They added only those vertices to the $V_{Extension}$ set that have two properties: Their label must be larger than that of v and they must only be neighboured to the newly added vertex w but not to a vertex already in a vertex already in $V_{Subgraph}$, that is, they must be in the exclusive neighbourhood of w with respect to $V_{Subgraph}$. Rand-ESU is a sampling algorithm to ESU (Wernicke, 2006) and this method has a lot of advantages, this kind of sampling algorithm is unbiased and easy to implement but there is a problem to solve. Rand-ESU gives each vertex in one layer the same probability value. For example vertex v and vertex u are in the same layer, the child node number of the sub-tree with v as its root is much higher than the sub-tree with u as its root. Because of the same probability, which vertex is chosen will affect the accuracy of the result. The analysis on the deficiency of Rand-ESU is showing in the Fig. 1 from the original Fig. 4 in the research of Wernicke (2006).

As show in Fig. 1a, the red part selects four vertexes with larger degree and larger $V_{Extension}$ in the first layer: 1, 2, 3 and 4, then continue to extend the sub-tree until get the size-3 subgraphs. There were 15 subgraphs digged out in Fig. 1a, it means that the sampling coverage to 93.75%. Figure 1b is on the contrary, the red part selects also four vertexes in the first layer: 3, 4, 5 and 6 but three of them have empty extension set; there is no way to expand the size of subgraph. At last, even by the sub-tree with vertex 3 as the root has been always selected, the sampling rate is only 6.25%. According to the example, each current subgraph has its own extension set with different size, so the randomness of the Rand-ESU sampling algorithm will greatly influence the accuracy of finding motifs. Most of the PPI networks are sparse and complex networks, different proteins involved in building different number of interaction edges. For example, most of the key proteins have larger degrees (Jeong *et al*., 2001), module areas are more dense to the other area in the whole network (Yu *et al*., 2010). Based on such consideration, a new sampling algorithm based on ESU has been proposed.

The proposed sampling algorithm is based on the adaptivity of the extension set of the subgraph called Adapt Rand ESU (AS-ESU). The basic idea is that when it needs to expand a new node, the larger the extension set of current subgraph is, the larger the sampling probability value will be given. Set a sampling probability P_k for every node for every node when add a new vertex in the current subgraph:

$$\left\{\begin{array}{l} P_k = \dfrac{|V_{Extension}|}{V_{max}[k]}, |V_{Extension}| \neq 0 \\ P_k = \dfrac{1}{V_{max}[k]}, |V_{Extension}| = 0 \end{array}\right\} \quad (1)$$

where, $|V_{Extension}|$ is the number of vertexes in the extension set of current subgraph i, $V_{max}[k]$ is the relative maximum value of the extension set size in the whole layer k. In order to obtain the bare maximum value of the whole layer,

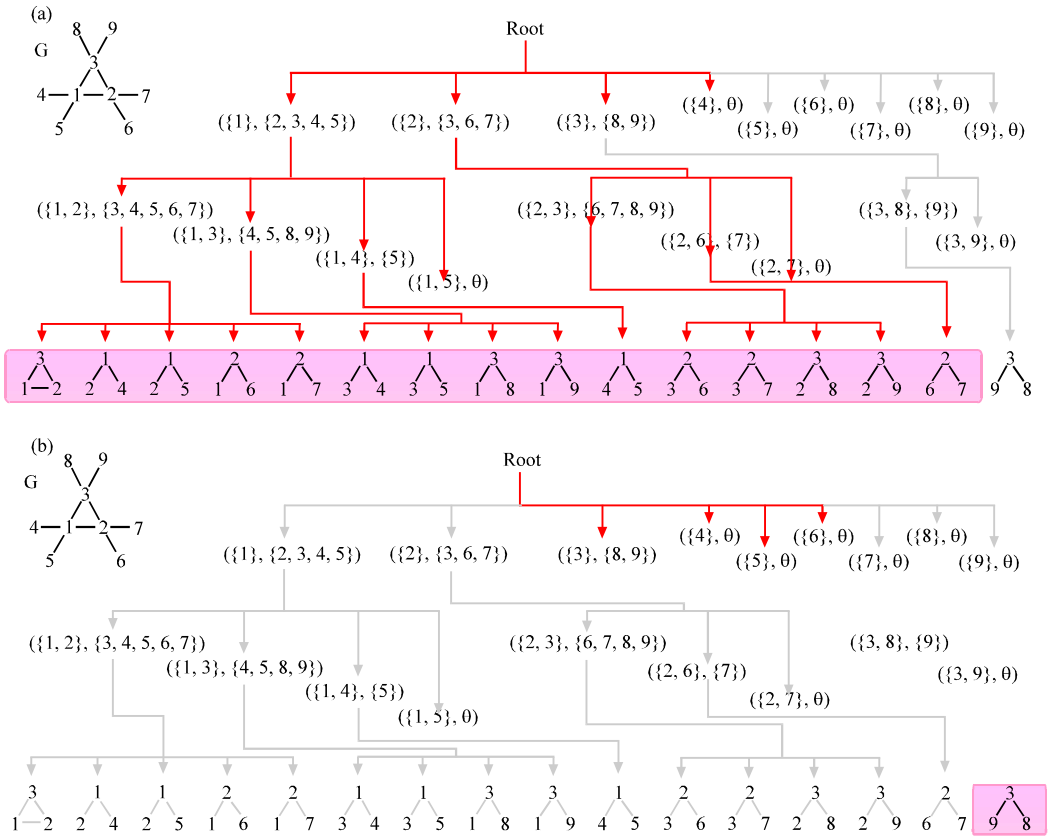


Fig. 1(a-b): Rand-ESU (Rand-enumerate subgraph,) sampling; (a) Select the vertexes whose $V_{Extension}$ is larger in the same layer and (b) Select the vertexes whose $V_{Extension}$ is smaller in the same layer

```
Algorithm: Adaptivity sampling-Enmuerate subgraphs (G, K) (AS-ESu)
Input: A grpah G = (V, E) and an integer 1≤k≤|V|
Output sampling size-k subgraphs in G.
01      for each vertex v∈V do
02      V_extension ←{u∈N({v}): u>v}
03      Call: Adaptivity Sampling Extend Subgraph ({v}, V_extension, v, 1)
04      return
Adaptivity Sampling Extend Subgraph (V_subgraph, V_extension, v, P)
AS1     if |V_subgraph| = k then output G [V_subgraph], P and return
AS2     set k' = |V_subgraph|, V_max [k'] = 1
AS3     if V_max [k']<|V_extension| then V_max (k') = |V_extension|
AS4     get a rendom R from [1, V_max[k']]
AS5     if R>|V_extension| then return
AS6     if |V_extension|>0 then P_k' = |V_extension|/V_max (K')
AS7     else P_k' = 1/V_max [k']
AS8     while V_extension≠ϕ do
AS9     Remove an arbitrarily chosen w from V_extension
AS10    V'_extension←V_extension∪ {u∈N_excl (w, V_subgraph): u>v}
AS11    call Adaptivity Sampling Extend Subgraph (V_subgraph∪{w}, V'_extension, v, P*P_k')
AS12    return
```

Fig. 2: Pseudocode for the new subgraph sampling algorithm, AS-ESU: Adaptivity sampling-enumerate subgraph

the algorithm need to calculate all the extension set size of every node in this layer. For this, each layer must get the denominator after massive calculation; this is against the purpose of designing a fast and effective algorithm. So the relative maximum value is selected rather than the bare maximum value in this study. Pseudocode of the new algorithm is shown in Fig. 2.

Subgraphs matching alignment: The motifs detection algorithm is put forward based on the definition of probability motifs proposed by Berg and Lassig (2004), so the next step needs to find a group of similar subgraphs, not completely isomorphism ones, their structural similarity between each other and they can have small differences. In the real biological theories, when the organisms being in constant evolution, motifs have also undergone a certain structure variations, which reflects the dynamic evolution of biological networks. Calculating the average value of a group of similar subgraphs' adjacency matrix to get the adjacency matrix of probability motif, the key step is how to compare different subgraphs (Qin and Gao, 2012), so there will need to introduce a judgment mechanism: subgraphs comparison algorithm.

A graph alignment is defined by a set of several subgraphs and a specific order of the vertexes in each subgraph; this joint order is again denoted by λ. Subgraph vertex matching is the first procedure to ensure λ. Vertex invariants are some inherent properties of the vertexes that do not change across mappings (Riaz *et al*., 2005). For simplicity, the step assumes here that the subgraphs are of the same size-k and detects probability motifs of different size separately.

The most convenient method is sorting the vertexes by their degree, from high to low, forming a one-one mapping. But different proteins have different function or other biological properties and motif play an important role in biological evolution. In this study, the subgraph matching algorithm complies with these rules: protein name matching first, then degree matching for the remaining vertexes.

For example in Fig. 3, G^{α} and G^{β} are two subgraph of size-4 and the purpose is getting the join vertex order of these two subgraphs by the matching rules. Protein A and protein B are both in the two subgraphs, so let them get matching first: {1, 2} in G^{α} and {2, 4} in G^{β}. For step 2, sorting the residual vertex according to degree from high to low in each subgraph, so there gets {3, 4} in G^{α} and {3, 1} in G^{β}. Finally, the vertex order is ensured.

Probability motifs classification: C^{α} is the adjacency matrix of subgraphs G^{α}, for any two aligned subgraphs G^{α} and G^{β}, the pairwise mismatch point is defined as follow (Berg and Lassig, 2004):

$$\text{Mismatch}\ (C^{\alpha}, C^{\beta}) = \sum_{i,j=1}^{n}[c_{ij}^{\alpha}(1-c_{ij}^{\beta}) + (1-c_{ij}^{\alpha})c_{ij}^{\beta}] \quad (2)$$

The mismatch point is 0 if and only if the matrices C^{α} and C^{β} are equal and is positive otherwise. It can be considered as a Hamming distance for aligned subgraphs. The lower the mismatch point is, the similar the subgraphs are and they are more likely to constitute a probability motif.

In order to derive probability motifs only from similar subgraphs and discover several motifs for a given size, from this study, a classification algorithm is used to find similar subgraphs and a probability motif for each cluster is evaluated. First, choose one subgraph G^1 initially as a cluster by itself. Then, calculate the mismatch point between G^1 and G^2, if the mismatch point is less than the mismatch threshold M_0, G^2 is joined to cluster 1 with G^1, otherwise, G^2 creates a new cluster. When G^n is going to

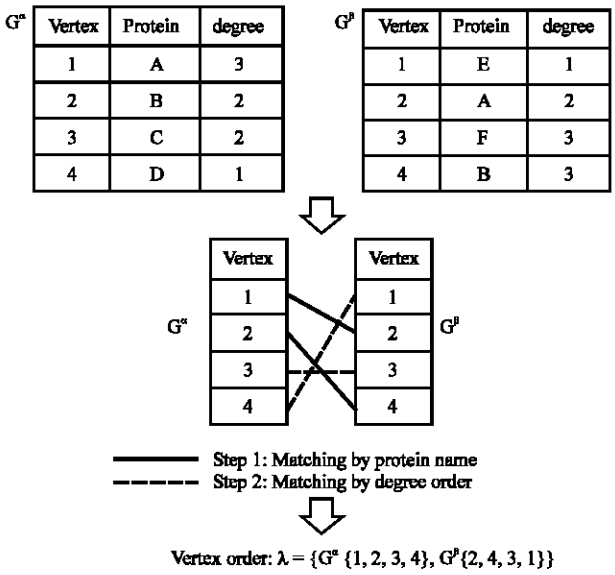


Fig. 3: Subgraph vertexes matching rules

Algorithm: Probability motifs classification of size-k
Input: A set of subgraphs $S = \{G^1, G^2, \ldots, G^n\}$ and mismatch threshold M_0
Output: A set of probability motifs $S' = \{CM^1, CM^2, \ldots, CM^m\}$

```
initialization: Cluster[1] = {G^1}, ClusterRepresentative[1] = G^1
for each G^m ∈ S, (m = 2,3, ..., n) do
    set Mismatch_min = Mismatch(G^m, G^1)
    set N = |ClusterPepresentative |
    for each ClusterPepresentative[i] ∈ ClusterPepresentative
        set M = Mismatch(G^m, ClusterPepresentativ[i])
        if M < Mismatch_min then Mismatch_min = M, set class = I
    if Mismatch_min ≤ M_0 then add G^m to Cluster[class]
        else ClusterRepresentative[N+1] = G^m, Cluster[N+1] = {G^m}
for each Cluster[i] ∈ Cluster do
    CM^i = CalculateMotifMatrix (Cluster[i])    // the average of the adjacency matrix
return the set of probability motifs S'
```

Fig. 4: Pseudocode for probability motifs classification algorithm

find its classification, compare G^n to each cluster to find the cluster M who has a minimum mismatch point with G^n. If the minimum is less than M_0, then put G^n in cluster M; otherwise G^n creates a new cluster. The pseudocode for probability motifs classification algorithm is showing as Fig. 4.

When each subgraph for a given size is divided into an exact classification as the Pseudocode showing, the probability motif CM, which is families of each classification, can be calculated by the set of n similar subgraphs $\{G^1, G^2, \ldots, G^n\}$:

$$CM = \frac{1}{n}\sum \text{Matrix}\{G^1, G^2, \cdots, G^n\} \quad (3)$$

RESULTS AND DISCUSSION

To evaluate the performance of this study, three PPI networks are used as test data: Saccharomyces cerevisiae PPI data (S-DIP) and yeast core PPI data (Core-DIP-20120518CR version) downloaded from DIP database (http://dip.doe-mbi.ucla.edu/dip/Main.cgi), UETZ data from Uetz *et al.* (2000). According to the three PPI networks, Table 1 shows the subgraph number of each size and that is the complete subgraph number without any sampling operation. The mining result showing in Table 1 is exactly the same with FANMOD method (Wernicke and Rasche, 2006). That mean the realization of the basic ESU code in this study is correct.

From Table 1 that complete subgraph set has a huge number with the increasing of size-k, it is difficult to continue follow-up calculation, so there must be a sampling mechanism that called AS-ESU based on improving Rand-ESU. The new algorithm makes sampling contingency smaller, guarantees the sampling rules covering more vertexes with higher degree and covering the relatively dense network area. Keep the same parameters, run AS-ESU and Rand-ESU twice to get the different sampling results in Table 2. The results of Rand-ESU come from FANMOD, each layer probability value equals to 0.5. From Table 2, two AS-ESU results of

Table 1: The number of subgraphs in different PPI databases

Database	UETZ	Core-DIP	S-DIP
Vertex number	1004	2191	4746
Edge number	957	4290	15166
Size-3	2377	31237	355412
Size-4	10510	343124	15509802
Size-5	56724	4364749	823897272
Size-6	337824	59103782	-
Size-7	2143248	832186474	-

UETZ: Database from Uetz *et al.* (2000), Core-DIP: Yeast core PPI database, S-DIP: Saccharomyces cerevisiae PPI database, PPI: Protein-protein interaction network

Table 2: AS-ESU and Rand-ESU sampling results comparison

	UETZ		Core-DIP		S-DIP	
Size	AS-ESU	Rand-ESU	AS-ESU	Rand-ESU	AS-ESU	Rand-ESU
3	492	203	5857	3410	33794	25465
	491	346	4223	4253	40675	56165
4	1198	414	32857	25195	1036677	1230671
	1372	606	35163	19328	843797	381629
5	2394	1125	184321	98586	25364934	18270317
	2645	1999	222703	185674	24178701	33692546
6	10538	1574	1037336	709618	-	-
	11184	4576	1053846	1149213		

UETZ: database from Uetz *et al.* (2000), Core-DIP: Yeast core PPI database S-DIP: Saccharomyces cerevisiae, PPI database, AS-ESU: Adaptivity sampling-enumerate subgraph, Rand-ESU: Rand-enumerate subgraph

size-3 in S-DIP database are 33794 and 40675, the Rand-ESU results are 25465 and 56165. The second result of AS-ESU is nearly 20% larger than the first time, while the second result of Rand-ESU is over 120% larger than the first time. Obviously, AS-ESU has better sampling stability and the same situation will appear in the other two databases. It means that the AS-ESU can better represent the original network topology property to improve the accuracy of motif detection.

The last work of the study, the goal is for all size-k subgraph classification according to the result of AS-ESU sampling algorithm running the 2nd time and following the rules: protein name matching first, vertex degree matching second. When size-k from low to high, the more edges will appear in the subgraphs, then there should allow more differences to ensure getting the most representative probability motifs for each size-k. So in parameters setting, mismatching threshold M_0 will increase with size-k and it is helpful to avoid the classification becoming too rough or too precise. Table 3 shows the classification result. For example, there are 4 probability motif classification of size-4 detected in UETZ database where the threshold is 2.With the increase of subgraph size, more classes will be found. Table 4 shows the top two probability motifs which have a larger percentage in each situation. The result shows the average probability value matrix and the topological graph of each probability motif. The larger the probability value of the matrix is, the denser the corresponding line in topological graph is. Berg and Lassig (2004) used the *E. coli* gene regulatory network with 424 nodes and 577 edges to find the probability motifs with size-5. Qin and Gao (2012) used the network whose size was similar to Berg and Lassig (2004) and they also only found the motifs with size-3 to size-5. PPI networks used in this study have bigger scale than other biological networks, based on the new subgraph sampling and matching algorithm, the study can find the probability motifs with bigger size than other motif detection methods used in smaller networks.

Table 3: Classification result in different PPI networks

Size	M_0	UETZ	Core-DIP	S-DIP
Size-3	0	2	2	2
Size-4	2	4	3	7
Size-5	3	4	12	-
Size-6	4	11	24	-

M_0: Subgraphs mismatching threshold, UETZ: Database from Uetz *et al.* (2000), Core-DIP: Yeast core PPI database, S-DIP: Saccharomyces cerevisiae PPI database, PPI: Protein-protein interaction network

Table 4: Top two probability motifs that have a larger percentage in each situation

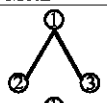
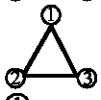
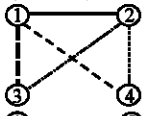
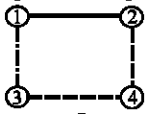
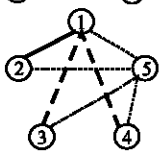
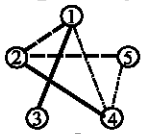
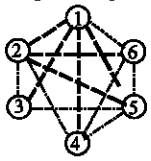
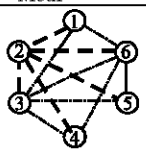
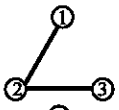
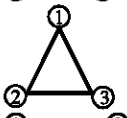
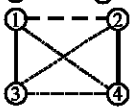
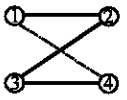
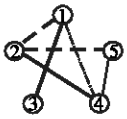
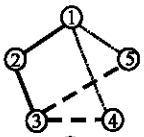
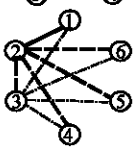
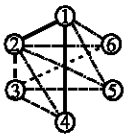
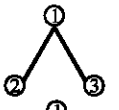
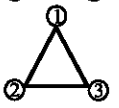
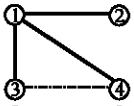
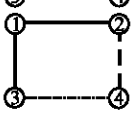
Database	Size	Matrix	Motif	p (%)
UETZ	3	$\begin{bmatrix}0&1&1\\1&0&0\\1&0&0\end{bmatrix}$		97.76
		$\begin{bmatrix}0&1&1\\1&0&1\\1&1&0\end{bmatrix}$		2.24
	4	$\begin{bmatrix}0&1&0.96&0.7\\1&0&0.07&0.31\\0.96&0.07&0&0\\0.7&0.31&0&0\end{bmatrix}$		77.55
		$\begin{bmatrix}0&1&0.04&0\\1&0&0&0.99\\0.04&0&0&0.98\\0&0.99&0.98&0\end{bmatrix}$		21.72
	5	$\begin{bmatrix}0&1&0.96&0.93&0.57\\1&0&0&0&0.45\\0.96&0&0&0&0.04\\0.93&0&0&0&0.07\\0.57&0.45&0.04&0.07&0\end{bmatrix}$		44.73
		$\begin{bmatrix}0&0.8&1&0.21&0\\0.8&0&0&1&0.91\\1&0&0&0&0\\0.21&1&0&0&0.1\\0&0.91&0&0.1&0\end{bmatrix}$		27.94
	6	$\begin{bmatrix}0&0.72&0.87&0.83&0.69&0.32\\0.72&0&0.04&0.01&0.64&0.67\\0.87&0.04&0&0&0.09&0\\0.83&0.01&0&0&0.16&0.01\\0.69&0.64&0.09&0.16&0&0.09\\0.32&0.67&0&0.01&0.09&0\end{bmatrix}$		28.11

Table 4: Continue

Database	Size	Matrix	Motif	p (%)
		$\begin{bmatrix} 0 & 0.97 & 0.13 & 0 & 0 & 0.03 \\ 0.97 & 0 & 0.87 & 0.95 & 0.96 & 0.98 \\ 0.13 & 0.87 & 0 & 0.01 & 0.01 & 0.02 \\ 0 & 0.95 & 0.01 & 0 & 0 & 0.06 \\ 0 & 0.96 & 0.01 & 0 & 0 & 0.03 \\ 0.03 & 0.98 & 0.02 & 0.06 & 0.03 & 0 \end{bmatrix}$		20.00
Core-DIP	3	$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$		94.67
		$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$		5.33
	4	$\begin{bmatrix} 0 & 0.99 & 1 & 0.04 \\ 0.99 & 0 & 0.37 & 1 \\ 1 & 0.37 & 0 & 0.01 \\ 0.04 & 1 & 0.01 & 0 \end{bmatrix}$		22.77
		$\begin{bmatrix} 0 & 1 & 0 & 0.01 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0.01 & 0 & 1 & 0 \end{bmatrix}$		15.66
	5	$\begin{bmatrix} 0 & 0.91 & 1 & 0.18 & 0 \\ 0.91 & 0 & 0 & 1 & 0.97 \\ 1 & 0 & 0 & 0 & 0 \\ 0.18 & 1 & 0 & 0 & 0.08 \\ 0 & 0.97 & 0 & 0.08 & 0 \end{bmatrix}$		15.95
		$\begin{bmatrix} 0 & 1 & 0 & 0.09 & 0.1 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0.92 & 0.9 \\ 0.09 & 0 & 0.92 & 0 & 0 \\ 0.1 & 0 & 0.9 & 0 & 0 \end{bmatrix}$		14.84
	6	$\begin{bmatrix} 0 & 1 & 0.12 & 0 & 0 & 0 \\ 1 & 0 & 0.99 & 0.83 & 0.86 & 0.88 \\ 0.12 & 0.99 & 0 & 0.16 & 0.14 & 0.13 \\ 0 & 0.83 & 0.16 & 0 & 0 & 0 \\ 0 & 0.86 & 0.14 & 0 & 0 & 0 \\ 0 & 0.88 & 0.13 & 0 & 0 & 0 \end{bmatrix}$		58.06
		$\begin{bmatrix} 0 & 1 & 0 & 0.02 & 0.12 & 0.14 \\ 1 & 0 & 0.99 & 0.03 & 0.07 & 0.07 \\ 0 & 0.99 & 0 & 0.99 & 0.81 & 0.78 \\ 0.02 & 0.03 & 0.99 & 0 & 0 & 0 \\ 0.12 & 0.07 & 0.81 & 0 & 0 & 0 \\ 0.14 & 0.07 & 0.78 & 0 & 0 & 0 \end{bmatrix}$		14.41
S-DIP	3	$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$		99.37
		$\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$		0.63
	4	$\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0.02 \\ 1 & 0 & 0.02 & 0 \end{bmatrix}$		84.08
		$\begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0.99 \\ 1 & 0 & 0 & 0.01 \\ 0 & 0.99 & 0.01 & 0 \end{bmatrix}$		8.23

UETZ: Database from Uetz *et al.* (2000), Core-DIP: Yeast core PPI database, S-DIP: Saccharomyces cerevisiae PPI database

CONCLUSION

To detect probability motifs, subgraph sampling and matching are necessary and important steps. In this study, a new algorithm is provided to calculate the probability value based on the adaptivity of the extension set of the subgraph; it is an increase of the traditional sampling algorithm Rand-ESU. According to the experimental results, the new algorithm has better stability and it is helpful to the final accuracy of motifs detection. The next, subgraph matching is based on protein type and vertex degree. That is very different from the traditional motif definition which only considers the topological characteristics. In the future, studies can do more experiments for similar subgraphs grouping; there must be many different clustering methods that can be used for probability motifs calculation. The future work should also pay attention to combine more biological information for motifs detection. The research concept of combining biological information and topological structure will be a new development direction on motifs detection and this will promote the research to whole biological network evolution.

ACKNOWLEDGMENTS

This study is supported by National Natural Science Foundation of China (Grant No. 61240046) and the key Natural Science Foundation of Hunan Province (Grant No. 13JJ2017).

REFERENCES

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

Alon, U., 2007. Network motifs: Theory and experimental approaches. Nat. Rev. Genet., 8: 450-461.

Berg, J. and M. Lassig, 2004. Local graph alignment and motif search in biological networks. PNAS, 101: 14689-14694.

Chen, J., W. Hsu, M.L. Lee and S.K. Ng, 2006. NeMoFinder: Dissecting genome-wide protein-protein interactions with meso-scale network motifs. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 20-23, 2006, Philadelphia, pp: 106-115.

Ciriello, G. and C. Guerra, 2008. A review on models and algorithms for motif discovery in protein-protein interaction networks. Briefings Funct. Genomics Proteomics, 7: 147-156.

Itzkovitz, S., R. Levitt, N. Kashtan, R. Milo, M. Itzkovitz and U. Alon, 2005. Coarse-graining and self-similarity of complex networks. Phys. Rev. E, 71: 1-10.

Jeong, H., S.P. Mason, A.L. Barabasi and Z.N. Oltvai, 2001. Lethality and centrality in protein networks. Nature, 411: 41-42.

Jiang, R., Z. Tu, T. Chen and F. Sun, 2006. Network motif identification in stochastic networks. PNAS, 103: 9404-9409.

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.

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

Qin, G.M. and L. Gao, 2012. An algorithm for network motif discovery in biological networks. Int. J. Data Min. Bioinf., 6: 1-16.

Riaz, K., M.S.H. Khiyal and M. Arshad, 2005. Matrix equality: An application of graph isomorphism. Inform. Technol. J., 4: 6-10.

Uetz, P., L. Giot, G. Cagney, T.A. Mansfield and R.S. Judson *et al*., 2000. A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. Nature, 403: 623-627.

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.

Xiang, X., D. Zhang, J. Qin and Y. Fu, 2010. Ant colony with genetic algorithm based on planar graph for multiple sequence alignment. Inform. Technol. J., 9: 274-281.

Yu, L., L. Gao and K. Li, 2010. A method based on local density and random walks for complexes detection in protein interaction networks. J. Bioinf. Comput. Biol., 8: 47-62.

Zou, Z., J. Li, H. Gao and S. Zhang, 2010. Mining frequent subgraph patterns from uncertain graph data. IEEE Trans. Knowl. Data Eng., 22: 1203-1218.