A Research on Network Similarity Search Algorithm for Biological Networks

SHEN Cong[1], DAI Xiao-peng[1], LI Dong-hui[1*]

[1]School of information science and technology of Hunan Agricultural university,410128 Changsha,China

Abstract. The biological network database presents exponential growth, how to find the target network accurately from the network database becomes the difficult problem. This paper proposes a new network similarity search algorithm, the similar network of Top k is calculated by two methods, the similar networks returned by the two algorithms are then filtered by overlap fractions, the weighted reordering algorithm is used to reorder the two sets of data, a precise set of similar network data sets is returned finally.In this paper, the accuracy of the query is judged by the comparison of the edge correctness (EC) value and the maximum public connection subgraph (LCCS) value of the returned sorted similar network data set, and compare query time with other algorithms.From the results, this algorithm is superior to other algorithms in query accuracy and query speed.

1 INTRODUCTION

The network is widely used in bioinformatics[1], chemical informatics[2], biomedicine[3], social network analysis[4], and other application fields[5]. High-throughput biological technology has been applied to produce large amounts of biological networks, such as compound structure network[6], biological pathways[7], transcription regulation network[8], protein-protein interaction networks[9], proteins-DNA interaction network[10]. It is difficult to find the target network in a large number of biological networks, and researchers have developed different kinds of Internet search technology: C-tree[11] is the indexing technology of k-nn query based on network editing distance. GString[12] is a semantic approach; GraphGrepSX[13] is an index subgraph similar search method based on suffix tree structure; SIGMA[14] is a collection based NSS method; RINQ[15] is a reference based index query method; NeMa[16] is a subgraph search method of a community; MAGE[17] is a pattern matching system that supports a random walk based network (RWR) algorithm; REFBSS[18] redefined RINQ's improvement. However, the above algorithm has limited query network, the query return value is empty, the query time is too long, and the query precision is not high enough etc question. Therefore, in view of the deficiency existing in the above algorithm, proposed a new algorithm, the algorithm by combining the two similarity search of Top k network to achieve the similar network the improvement of accuracy and less time for the query.

2 DEFINITIONS AND METHODS

2.1 Network Database and Query Network

A network can be regarded as a directed graph N = (V, E), V represents the point in the graph, E represents the edge in the graph, and the network database is the data center used to store the biological network. The network database can be expressed as $D=\{N_1,N_2,...N_n\}$, which contains n networks, where Ni represents the ith network in the network database. The query network is expressed as $T=\{Q_1,Q_2...Q_q\}$, where Q_j represents the jth network in the query network.

2.2 Subnet

2.2.1 subnet definitions

If there is a network N' meet: The point V' in N' is a subset of the point V in N, The edge E' in N' is a subset of edge E in N, The network N'(N',V') is the subnet of network N(V,E), which can be abbreviated as N' is the subnet of N. For biological networks, the nodes in the network are biological molecules and the edges are intermolecular interactions. The eigenvectors of two nodes, three nodes and four nodes can be represented as $Sub_2N=[Sub_2N_1, Sub_2N_2,\cdots,Sub_2N_n]^T$, $Sub_3N=[Sub_3N_1, Sub_3N_2,...,Sub_3N_n]^T$, and $Sub_4N=[Sub_4N_1,Sub_4N_2,\cdots, Sub_4N_n]^T$. The standardization of corresponding subnet frequency (See fig 2.2.3 frequency calculation of nodes)

for Sub2Ni=[f(i,1),f(i,2)],Sub3Ni=[f(i,1),f(i,2),⋯,f(i,13)], Sub4Ni=[f(i,1),f(i,2),⋯,f(i,199)].

2.2.2 divide the subnet

* Corresponding author: author@e-mail.org

Subnetting part, this article uses the MFinder [19] algorithm, there are two kinds of connections between nodes and nodes, unidirectional and bidirectional connections, so for two nodes graphs there are two types, as shown in figure 1(a), there are 13 types of three nodes graphs, as shown in figure 1(b), four nodes graphs type has 199 kinds, with the increase of graph nodes, the type of subgraph takes on the form of exponential growth, therefore, in this paper, the query and the target network subnet partition only two nodes subnet, 3 subnet and 4 node subnet.

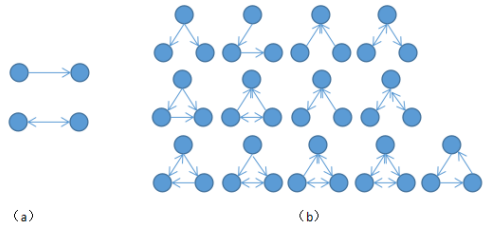

Fig1. The types of the two nodes and the three nodes.

2.2.3 subnet frequency calculation

Subnet frequency calculation is a relatively complex process. Assuming that n nodes subgraph, are connected in a node on the edge of n-1 the maximum, to calculate the probability P of a child graph, we need to consider likely to n-1 the edge, the probability of the emergence of the subgraph is equal to the graph of each side in a probability sum, computation formula is as follows:

$$P = \sum_{\sigma \in S_m} \prod_{E_j \in \sigma} \Pr[E_j = e_j \mid (E_1,...,E_{j-1}) = (e_1,...,e_{j-1})].$$

s_m is the set of all sequence permutations of all (n-1) edges, and E_j is the jth edge in the n-1 dimension. In the following figure, we calculate the probability of occurrence of subgraph {1,2,3} and the probability of occurrence of subgraph {4,5,6} :

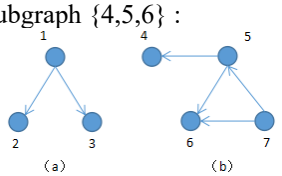

Figure (a) the probability of the occurrence of {1,2,3}	Figure (b) the probability of the occurrence of {4,5,6}
1 Pick first(1,2)，Pr=1/E=1/6. Then (1,3)，Pr=1. So Pr[(1,2),(1,3)]=1/6*1=1/6. 2 Pick first(1,3)，Pr=1/E=1/6.	1 Pick first(5,4)，Pr=1/E=1/6. Then (5,6)，Pr=1/2. So Pr[(5,4),(5,6)]=1/6*1/2=1/12. 2 Pick first(5,6)，Pr=1/E=1/6.

Fig 2 Analysis of subgraph frequency calculation.

By the algorithm, Although the {1, 2, 3} in the figure a and the {4 5 6} in the figure are homogeneous subgraph, there are different subgraph probabilities in different graphs, probability of subgraph for subsequent similar network query to provide powerful guarantee.

2.3 Cosine Similarity

Assuming that SubQ and SubN represent the k node subnet of network Q and network N respectively, then the cosine similarity calculation of network Q and network N is shown in formula:

$$\cos(Q,N) = \frac{\sum_{i=1}^{n} SubQ_i \cdot SubN_i}{\sqrt{\sum_{i=1}^{n} (SubQ_i)^2}\sqrt{\sum_{i=1}^{n} (SubN_i)^2}}$$

Q and N of cosine similarity calculation is done by its corresponding subnet, N is the number of subnets, taking an example of 2 nodes subgraph, there are only two possible ways to connect the two idea graphs, so n takes 2, the probability of target network and the query network of subnet A are: 2/3 and 1;the probability of target network and the query network of subnet B are: 1/3 and 0. The cosine similarity of the two networks can be calculated as:

$$\frac{\frac{2}{3}\times 1 + \frac{1}{3}\times 0}{\sqrt{(\frac{2}{3})^2 + (\frac{1}{3})^2}\sqrt{1^2 + 0^2}} = 0.8944$$

2.4 Network Alignment Quality Index

An important indicator for measuring network similarity is network alignment(NA). and network matching is divided into local networks than (local network alignment LNA) and global network than (global network alignment GNA), local network than the main concern is the biological information, such as correlation function consistency and biology; While global network comparison focuses on biological information and topology information, the following two indicators are the two most common methods used to judge topological similarity.

2.4.1 edge correctness(EC)

EC is the percentage for edges in network N_i that are aligned to network N_j, so EC worth the value range of [0,1], the two network N_i= (V_i, E_i) and Nj= (V_j, E_j), the contrast of two networks can be expressed as injective function f: $V(N_i)$->$V(N_j)$. The calculation formula of EC is defined as follows:

$$EC = \frac{|\{(u,v) \in E_i : (f(u), f(v)) \in E_j\}|}{|E_i|}$$

2.4.2 largest common connected subgraph(LCCS)

LCCS is the number of edges in the largest connected subgraph for the first network that is isomorphic to a subgraph of the second network. The value of LCCS is different from that of EC value, and the value range of LCCS is [0,$|E_i|$], and $|E_i|$ refers to the total number of edges in the first network. For the two networks with the same EC value, the network with large LCCS value is higher, while the larger the LCCS value is, the more dense the network is.

3 NETWORK SIMILARITY SEARCH ALGORITHM

3.1 Algorithm Overall Flow

Network similarity algorithm mainly divides into three parts, as shown in figure 3, the first part is the cosine similarity calculation, by calculating the cosine similarity between query and target network, returns the similarity ranking Top k network collection D_1; In the second part, the comparison parameters between the query network and the target network are calculated by EC and LCCS, and the network set D_2 is returned by the comparison parameter ranking Top k '(k' =k).The third part is divided into two steps: the first step is obtained by cosine similarity was calculated by the overlap of the Top k network and by EC value and LCCS is worth to the Top k' network of overlapping ratio, if more than the threshold value of π, go to the next step,or directly to the end, the query fails; The second step is to set weights for Top k network and Top k' network, for the two methods have been the former Top k network comprehensive ranking again, finally, returns a similar ranking Top k_2 network collection.

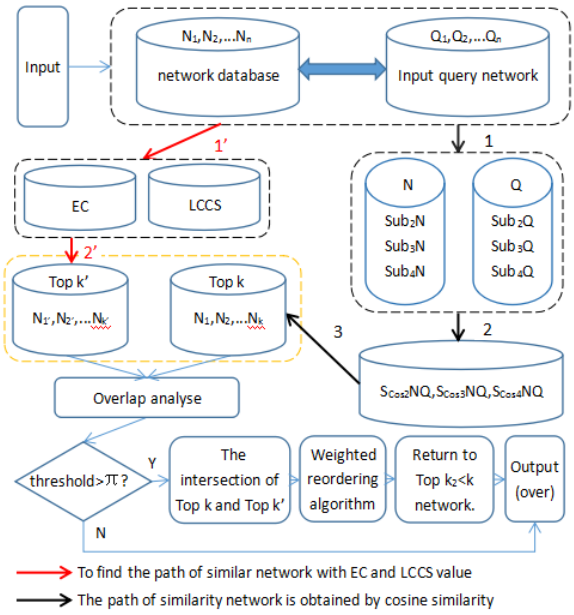

Fig 3 overall frame diagram.

3.2 Network Similarity Search

3.2.1 The cosine similarity gets the Top k network

The cosine similarity computing similarity Top k network in the process, first of all to the query and the target network partition subnet, because with the increase of subnet number of nodes, subnet number type can present the form of exponential growth, therefore, in this paper, in consideration of time complexity and computational complexity, in terms of the selection of subgraph, using only the section nodes 2, 3 and 4 nodes figure, after subnetting, the subnet is used to calculate cosine similarity between query and target network, and then based on the cosine similarity value as the query and the target network similarity criterion is an important standard, return to the former Top k similarity network, represented as $D_1 = \{N_1, N_2...N_k\}$.

3.2.2 EC value and the Top k network obtained by LCCS

EC value and LCCS is used to measure an important indicator of network than EC and LCCS value to a certain extent, reflects the degree of similarity between the network, so this article use this way as the second measurement network of similarity between the reference index of the first network and the query target network computing EC value, because the EC value as a percentage, in most of the small-scale network, as a result of the limitation of calculation accuracy, presents the difference is small, can't accurate judgment to the similarity between the network, in this case, in the case of small EC value differences, this paper users the second measurement value of LCCS supplement for EC value to calculate again LCCS little difference value, and then according to the EC value and LCCS worth comprehensive evaluation standard, returns the Top k' similarity similarity network, remember to $D_2=\{N_{1'},N_{2'}...N_{k'}\}$, where k'=k, is just to distinguish the data from D_2 and D_1.

3.2.3 Overlap and weighted reordering algorithm.

(1)Overlap coefficient

Due to the difference between D_1 and D_2 in data set, in order to judge the difference of D_1 and D_2, the Overlap is used to calculate the difference between the two data sets.

$$Overlap = \frac{|D_1 \cap D_2|}{|D_1 \cup D_2|}$$

We can clearly see that in the above equation. the more the same network between D_1 and D_2, the greater the value of Overlap, the two methods got similar networks overlaps the Top k is higher, the higher the accuracy, the Overlap of the peak can reach 1, namely two ways to get the Top k network exactly the same. If Overlap value is less than a threshold tends to zero, (take the experience value π =0.1 in this paper)we think that the network query failed, at least in two ways that there is an obvious error, one way to abandon the query.

(2)Weighted reordering algorithm

Two algorithms for the Top k similar network under the condition of satisfying Overlap, need to get the Top k similar in the two methods integrating network, and get a new sort, this article put forward the scheme of setting weights, similar to reorder, Top k network $D = \{N_1,N_2...N_k\}$ The weight setting formula of each network Ni is:

$$w_{N_i} = 1 - \frac{i}{k}$$

From the above formula can know, when the Top k, the lower the ranking network, the smaller the weight, to the Top of the network, the greater the weight, For the

intersection of D_1 and D_2, we pass and reorder the back number:

$$Order_i = Order_{D_1} \times (1 - \frac{i_{D_1}}{k_{D_1}}) + Order_{D_2} \times (1 - \frac{i_{D_2}}{k_{D_2}})$$

iD1 represents the ranking of the ith in the new sequence in the network set D_1, i_{D2} is the same, k_{D1} represents the size of k in D_1 network, and k_{D2} is the same. The Orderi value obtained by this algorithm is reordered from small to large, and a new sequence of similar networks is considered.

4 THE SIMULATION RESULTS

4.1 The Data Source

This article users four real data sets. First comes from the NCI/NIH AIDS antiviral drug screening data (http://dtp.cancer. gov), the molecular structure of the data set, the other three data sets are biological pathways data sets, can be downloaded from WikiPathways website, one of which is the Bos Taurus path data sets, the other two data sets are Homo Sapiens pathway, Homo Sapiens I and II in training network model is different, Homo Sapiens I was randomly selected from the data set while training the query network, Homo Sapiens II can only train up to 30 data sets when training the network. The network dimensions of the vertices and edges used in the experiment are listed in table 1.

Table 1 the number of four true data points and edges.

dataset	network database			query network		
	network	point	edge	network	point	edge
AIDS	500	3-176	3-182	30	7-24	6-25
Bos Taurus	200	8-360	7-314	30	5-359	5-371
Homo Sapiens I	609	5-855	5-648	30	7-440	6-388
Homo Sapiens II	609	5-347	5-273	30	303-855	291-648

4.2 Results analysis

4.2.1 Return the results of the EC and LCCS values

For the result of the final return, the network computing EC and LCCS values in the network and network database are shown in figure 4 and figure 5. EC value calculation, for data collection of AIDS, The values of the two nodes and the c-tree algorithm are not much different, and the EC values of the 3 nodes and 4 nodes are significantly higher than the c-tree algorithm. and Bos Taurus data sets, Homo Sapiens I data sets and Homo Sapiens II data sets, both nodes 2, 3, and 4 nodes, EC values were significantly higher than C-tree algorithm. LCCS value calculation, for data set AIDS and Homo Sapiens I, the LCCS value of 2 nodes is similar to that of c-tree algorithm, The LCCS value of 2 nodes in Homo Sapiens II is slightly lower than the c-tree algorithm, and other data set the remaining quarter idea figure of LCCS values are higher than c-tree algorithm. On the whole,whether the EC values as a measure, or the LCCS value as a measure, the algorithm of similar web search performance is much better than c-tree algorithm.

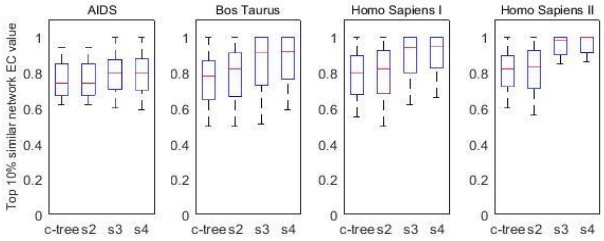

Fig 4 Result set Top 10% similar network EC value

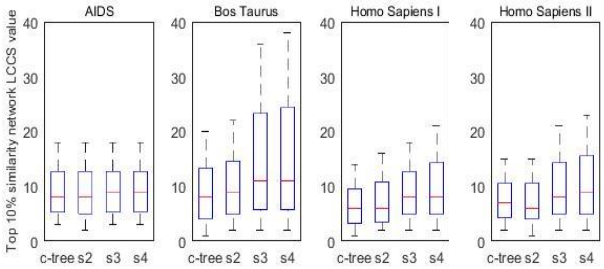

Fig 5 Result set Top 10% similar network LCCS value

4.2.2 Average query time

This study analyses the query time, the results of the analysis as shown in figure 6, two, three, four nodes and c-tree algorithms in four data sets are compared respectively, the results of the search results are better than that of the c-tree algorithm, and as a result, this algorithm not only improve the precision of the similar web search, and to a certain extent, reduce the network similarity search of time.

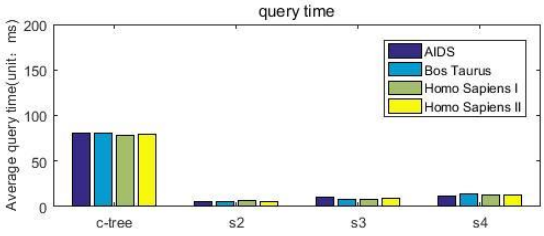

Fig 6 Query time comparison analysis chart

5 CONCLUSION

This paper, using cosine similarity and similar network and EC value, LCCS combination of the Top two k similar sequence set network, according to the two sequences set with Overlap judgment, then through reverse weighted weight sorting algorithm on two Top k get a collection of sequence data integration. This algorithm also performs a performance comparison with several other algorithms, which can be concluded as follows: (1)the algorithm improves the accuracy of network search; (2)optimized the algorithm and reduced the query time; (3)avoid the situation where the traditional method is limited by the query condition, and the return value of the query network is empty, because

the algorithm returns the network set of the previous k in the similarity degree.

FUNDING

The work described in this paper was partially supported by national key research and development project (Project No: SQ2017YFNC050022-06),Hunan education department scientific research project(Project No: 17K044 ; 17A092).

References

1. Von M C, Krause R, Snel B, et al. Comparative assessment of large-scale data sets of protein-protein interactions.[J]. Nature, 2002, 417(6887):399.
2. Rual J F, Venkatesan K, Hao T, et al. Towards a proteome-scale map of the human protein¿protein interaction network[J]. Nature, 2005, 437(7062):1173-8.
3. Szklarczyk D, Franceschini A, Kuhn M, et al. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored[J]. Nucleic Acids Research, 2011, 39(Database issue):561-8.
4. Leskovec J, Sosič R. SNAP: A General Purpose Network Analysis and Graph Mining Library.[J]. Acm Transactions on Intelligent Systems & Technology, 2016, 8(1):1.
5. Robinson I, Webber J, Eifrem E. Graph Databases: New Opportunities for Connected Data[M]. O'Reilly Media, Inc. 2015.
6. Willett P, Barnard J M, Downs G M. Chemical Similarity Searching[J]. J.chem.inf.comput.sci, 1998, 38(6):983--996.
7. Kanehisa,M. and Goto,S.Kegg: kyoto encyclopedia of genes and gen-omes. Nucleic Acids Res.2000, 28, 27–30.
8. Raymond,J.W. et al.Rascal: Calculation of graph similarity using max-imum common edge subgraphs. Comput. J.2002, 45, 631–644.
9. Panni,S. and Rombo,S.E. Searching for repetitions in biological networks: methods, resources and tools. Brief. Bioinf.2015, 16, 118–136.
10. Xu K, Schadt E E, Pollard K S, et al. Genomic and network patterns of schizophrenia genetic variation in human evolutionary accelerated regions.[J]. Molecular Biology & Evolution, 2015, 32(5):1148-60.
11. He H, Singh A K. Closure-Tree: An Index Structure for Graph Queries[C]// International Conference on Data Engineering. IEEE, 2006:38.
12. Jiang H, Wang H, Yu P S, et al. GString: A Novel Approach for Efficient Search in Graph Databases[C]// IEEE, International Conference on Data Engineering. IEEE, 2007:566-575.
13. Bonnici V, Ferro A, Giugno R, et al. Enhancing Graph Database Indexing by Suffix Tree Structure[C]// Pattern Recognition in Bioinformatics -, Iapr International Conference, Prib 2010, Nijmegen, the Netherlands, September 22-24, 2010. Proceedings. DBLP, 2010:195-203.
14. MISAEL MONGIOVu00cc, RAFFAELE DI NATALE, ROSALBA GIUGNO, et al. SIGMA: A SET-COVER-BASED INEXACT GRAPH MATCHING ALGORITHM[J]. Journal of Bioinformatics & Computational Biology, 2010, 8(02):199-218.
15. Günhan G, Tamer K. RINQ: Reference-based Indexing for Network Queries[J]. Bioinformatics, 2011, 27(13):i149-i158.
16. Khan A, Wu Y, Aggarwal C C, et al. NeMa: fast graph search with label similarity[C]// International Conference on Very Large Data Bases. VLDB Endowment, 2013:181-192.
17. Pienta R, Tamersoy A, Tong H, et al. MAGE: Matching Approximate Patterns in Richly-Attributed Graphs[C]// IEEE International Conference on Big Data. IEEE, 2014:585-590.
18. Soylev A, Abul O. REFBSS: Reference based similarity search in biological network databases[C]// Computational Intelligence in Bioinformatics and Computational Biology. IEEE, 2015:1-8.
19. Kashtan N, Itzkovitz S, Milo R, et al. Efficient sampling algorithm for estimating subgraph concentrations and detecting network motifs[J]. Bioinformatics, 2004, 20(11):1746-1758.