An Efficient Drug-Target Interaction Mining Algorithm in Heterogeneous Biological Networks

Congcong Li[1(✉)], Jing Sun[1], Yun Xiong[1], and Guangyong Zheng[2]

[1] Shanghai Key Laboratory of Data Science, School of Computer Science, Fudan University, Shanghai 200433, People's Republic of China
{congconglil2, jingsun, yunx}@fudan.edu.cn
[2] CAS-MPG Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, People's Republic of China
zhenggy@sibs.ac.cn

Abstract. The identification of interactions between drugs and targets is a key area in drug research. Exploring targets can help identify potential side effects and toxicities for drugs, as well as new applications of existing drugs. Because of the enormous scale of biological dataset, most of the existing algorithms for drug-target mining are time-consuming. In this paper, we proposed an optimization algorithm called LSH-HeteSim to mine the drug-target interaction in heterogeneous biological networks, where the relationship between drugs and targets is various. It means drugs and targets are connected with complicated semantic path. In practice, the similarity measure used for semantic path is a path-dependent method, called HeteSim, which had been utilized in some previous studies of relevance search. Experiment results in real biological networks show that our algorithm can effectively predict drug-target interaction with the *AUC* measure achieving 0.943. Simultaneously, the running time of our algorithm is much less than the state-of-art methods.

Keywords: Drug target · Link prediction · Heterogeneous biological networks · Meta-path · Similarity measure

1 Introduction

The key issue of modern drug development is to recognize drug targets. Drug targets are binding sites of drug and biological macromolecules regulated by the drug, such as receptors, enzymes, ion channels, transporters, genes and the like [1]. As the basis of drug discovery and designing, prediction of drug-target interactions is an important issue in the field of biological research field [2].

In traditional biological studies, whether there is a link between the drug and target gene is inferred from biology experiment results which have a long and costly experimental period [3]. In recent years, many endeavors have been made in drug-target interaction prediction, for accelerating drug targets finding, shortening research

W.-C. Peng et al. (Eds.): PAKDD 2014 Workshops, LNAI 8643, pp. 65–76, 2014.
DOI: 10.1007/978-3-319-13186-3_7

cycle, and reducing development costs [4]. However, most of the endeavors only considered partly characteristics of drug biological networks, such as drug chemical characteristics, or local link information. A biological network is a complex heterogeneous network, which either contains multiple types of objects or multiple types of links [7, 8]. Commonly, in a biological network, the relationship between drug and target is heterogeneous [9, 11], where two objects connected via different paths have different meaning [10]. Hence, for drug target similarity measurement, the semantic paths must be taken into account. Besides, the dataset of biological network is general large, and similarity search algorithms frequently cost a long time. To solve the problem mentioned above, we propose a new drug-target interaction mining algorithm, for similarity path search in heterogeneous biological networks. The new algorithm is named LSH-HeteSim, which is based on the locality sensitive hash method (LSH) [17]. The HeteSim similarity measure method has been utilized in some relevance search problem of social network [10], and it is also employed in our study. Experiments results show that the algorithm we proposed can predict missing links between drugs and targets and identify drug-target interaction with a fairly high accuracy (the *AUC* measure achieve 0.943). Specially, the running time of our algorithm is much less than the state-of-art methods.

The rest of this paper is organized as follows. In Sect. 2, we review some related work about link prediction and existing predictors for drug-target interaction finding. Next, in Sect. 3, we provide detail information of the relevance search algorithm HeteSim and our new drug-target interaction mining algorithm LSH-HeteSim. Then in Sect. 4, we perform two groups of experiments to prove the effectiveness and efficiency of our algorithm. Finally, in Sect. 5, we give discussion and conclusion of this study.

2 Related Work

2.1 Link Prediction

Link prediction aims at estimating the likelihood of the existence of a link between two nodes [12]. In some networks, especially biological networks such as PPI, metabolic networks and food webs, the discovery of links is costly in the laboratory [13]. Instead of blindly checking all possible links, predictions based on the observed links and focusing on those links which are most likely to exist can sharply reduce the experimental costs assuming that the prediction algorithm is accurate.

Node similarity based link prediction method can be roughly categorized into two types: feature based approaches and link based approaches. The feature based approaches measure the similarity of nodes based on their feature values, such as cosine similarity, Jaccard coefficient and Euclidean distance. The link based approaches measure the similarity of nodes based on their link structures in a network. The similarity measure (HeteSim) used in our proposed LSH-HeteSim algorithm is a link based method, especially, it takes the semantic paths into account.

2.2 Methods for Drug-Target Interaction Prediction

The prediction of drug-target interaction is an important research problem in the drug discovery filed. Traditional methods for drug target prediction are based on biological experiments. Due to the long test period and its high cost, there have been more and more drug target prediction methods by calculating. For example, Campillos and Monica proposed a method for identifying drug target genes by the similarity of side effects [4]. He Zhisong and Zhang Jian proposed a method based on the drug and biological characteristics of the functional group [5]. Chen Bin and Ying Ding used a statistical model called SLAP to measure the relationship between the drug and target gene in a biological network [6, 18]. As SLAP takes the heterogeneity of biological networks and the impacts of different semantic paths to the drug targets' similarities into consideration during its statistical computation, it achieves good prediction accuracy. However, due to its complicate computation, the experiments are conducted with sub-datasets on small scale. Once it applies to large scale dataset, the query task will be very time-consuming. Besides considering the semantic paths in heterogeneous biological networks, the optimized algorithm proposed in paper, called LSH-HeteSim, also adopt the LSH. It can reduce the running time sharply without losing the prediction accuracy. We will compare the prediction accuracy of these two algorithms on the same dataset in experimental section.

3 Drug-Target Interaction Mining in Heterogeneous Biological Networks

3.1 Heterogeneous Biological Networks

A heterogeneous network is a special type of network which either contains multiple types of objects or multiple types of links [10], while a heterogeneous biological network is composed of multiple biological objects. In many applications, the network object is no longer constituted with a simple type, but includes various types of objects and links, such as the gene regulatory networks. Here, we give a definition of heterogeneous biological networks as shown in Definition 1.

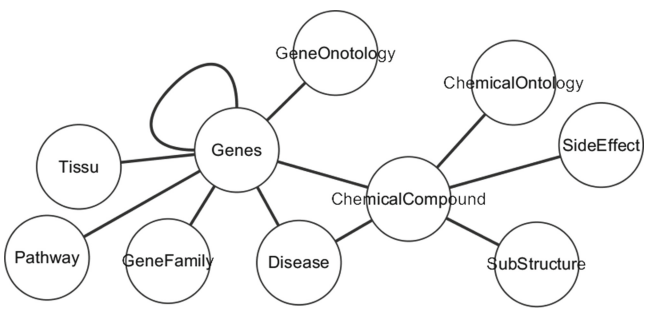

Fig. 1. Network schema of Slap

Definition 1. Heterogeneous biological Networks

Given a network schema $S_G = (A, R)$ which consists of a set of biological objects types $A = \{A\}$ and a set of relations $R = \{R\}$, an biological network is defined as a directed graph $G = (V, E)$ with an object type mapping function $\emptyset(v) \in A$ and a link type mapping function $\varphi(e) \in R$. Each object $v \in V$ belongs to one particular object type $\emptyset(v) \in A$, and each link $e \in E$ belongs to a particular relation $\varphi(e) \in R$. When the types of objects $|A| > 1$ or the types of relations $|R| > 1$, the network is called heterogeneous biological network. Otherwise it is a homogeneous biological network.

Figure 1 is the network schema of a heterogeneous biological network Slap [6]. Each circle represents a biological object type, such as drug, disease, gene, and etc. Each horizontal line represents a relation type, such as the treatment between drug and disease.

Definition 2. Meta-Path

A meta-path P is a path defined on the graph of network schema $T_G = (A, R)$, and is denoted in the form of $A_1 \xrightarrow{R_1} A_2 \xrightarrow{R_2} \ldots \xrightarrow{R_l} A_{l+1}$, which defines a composite relation $R = R_1 \circ R_2 \ldots \circ R_l$ between types A_i and A_{l+1}, where $\circ$ denotes the composition operator on relations.

Different from those in homogeneous network, the paths in heterogeneous network called meta-path have semantics, as Definition 2 defined, which make the relatedness between two objects different on different search paths. Taking the heterogeneous network in Fig. 2 for example, the relationship between the drug and the disease is treatment and being treated, and between genes and disease is the causing and caused by. Obviously, drug C_1 is not related to gene G_2 based on CDG path. It means that drugs can treat diseases caused by genes. However, drug C_1 is related to gene G_2 based on CDCDG path because of drug C_2 which can also treat disease D_2.

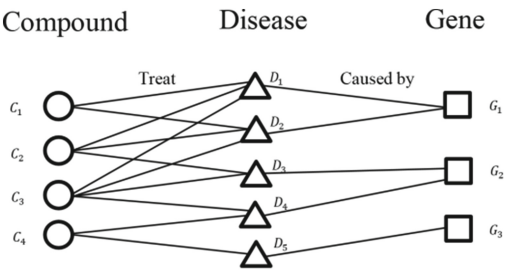

Fig. 2. A simple heterogeneous network example

3.2 Relevance Search

Shi Chuan and Kong Xiangnan have defined the relevance search problem in heterogeneous networks. They proposed an algorithm based meta-path a meta-path called HeteSim to measure the similarity between objects of different types [10]. In this paper, HeteSim was used to assess the interactions between drugs and targets in heterogeneous networks.

Definition 3. HeteSim [10]

$$HeteSim(s,t|R_1 \circ R_2 \circ \ldots \circ R_l) = \frac{1}{|O(s|R_1)||I(t|R_l)|} \sum_{i=1}^{|O(s|R_1)|} \sum_{j=1}^{|I(t|R_l)|} HeteSim(O_i(s|R_1), I_j(t|R_l)|R_2 \circ \ldots \circ R_{l-1}) \tag{1}$$

Given a meta-path $P = R_1 \circ R_2 \ldots \circ R_l$, HeteSim($s$, $t|P$) is the similarity between object s and t. Here, $O(s|R_1)$ is the out-neighbors of s based on the path P, and $I(s|R_l)$ is the in-neighbors of t based on P. The result it returns is a similarity value between 0 and 1. The larger HeteSim value is, the more similar the two objects are. The similarity of drug C_1 and gene G_1 was calculated as follows:

$$HeteSim = \frac{1}{|O(C_1|CD)||I(G_1|DG)|} \sum_{i=1}^{|O(C_1|CD)|} \sum_{j=1}^{|I(G_1|DG)|} HeteSim(O_i(C_1|CD), I_j(G_1|DG))$$

where $O(C_1|CD) = \{D_1, D_2\}, I(G_1|DG) = \{D_1, D_2\}$. So the *HeteSim* value of C_1 and G_1 is 0.5.

In order to facilitate the calculation, the formula in Definition 3 can be normalized as Eq. (2):

Definition 4. NormHeteSim [10]

$$NormHeteSim(x_i, x_j|P) = \frac{M_{P_L}(x_i,:)M_{P_R^{-1}}(x_j,:)}{\sqrt{|M_{P_L}(x_i,:)||M_{P_R^{-1}}(x_j,:)|}} \tag{2}$$

where M is the relation matrix defined as follows:

Definition 5. Relation Matrix

$$M = U_{A_1A_2} U_{A_2A_3} \ldots U_{A_{l-1}A_l} \tag{3}$$

where $U_{A_iA_j}$ is the adjacency matrix A_i and A_j.

In addition, $M_P(i, j)$ represents the number of path instances of meta-path $P = (A_1A_2 \ldots A_l)$ which start from object $x_i \in A_1$ to object $x_j \in A_l$, and $M_{P_L}(x_i,:)$ represents the feature vector of object x_i whose length is decided by the target object type A_l and $M_{P_R^{-1}}(b,:)$ represents the feature vector of object x_j, so the HeteSim value is the cosine similarity of the two feature vector. The path P_L and P_R in NormHeteSim definition is the decomposition of the path P from the middle position. That is, $P_L = (A_1A_2 \ldots A_{mid})$ and $P_R = (A_{\text{mid}}A_{\text{mid}+1} \ldots A_l)$.

In biological heterogeneous network, measuring the similarity of drug and target is suited to the relevance search problem definition. Therefore, HeteSimQuery is considered baseline algorithm to measure the similarity between drug and target as Algorithm 1 shows. For any given meta-path which starts with drug type and end with

target type, we can use HeteSimQuery to calculate the similarity of the query pairs (p, q). For example, if the similarity of one drug and gene pair under the path CGCDG (Compound-Gene-Compound-Disease-Gene) was queried, HeteSimQuery will return the similarity value of the query pair.

Algorithm 1. ***HeteSimQuery*** (p, q)

Input: data set D**, query object pair** (p, q)**, meta-path** P

Output: a similarity value of the query pair

1. **Separate** P **to** P_L **and** P_R
2. **Calculate the Relation Matrix** M_{P_L} **and** $M_{P_R^{-1}}$
3. **Generate Hash Vector of** p **and** p
4. **Calculate the similarity of** (p, q) **as Formula (2)**

Obviously, when the inputing meta-path $P = (A_1 A_2 \ldots A_l)$ of HeteSimQuery is symmetrical or the starting object type is the same as the end object type, that is A_1 and A_l are the same object type, such as CDC or CDGDC, HeteSimQuery can be used to measure the similarity between objects with the same types. Therefore, our baseline algorithm HeteSimQuery can be used to measure the similarity between objects of the same types as well as different types.

3.3 Locality Sensitive Hash

Locality Sensitive Hash (LSH) functions were introduced to solve the approximate nearest neighbor problem in high dimensional spaces [14]. It is designed in such a way that if two objects are close in the intended distance measure, the probability that they are hashed to the same value is high, and if they are far in the intended distance measure, the probability that they are hashed to the same value is low [15]. Here the meaning of 'close' and 'far' depend on the similarity measure used, and the exact formulation of LSH functions varies with the exact distance definition of the similarity measure. Nevertheless, all LSH functions should always comply with the locality sensitive hashing schema [16].

However, not every similarity measure has its corresponding LSH functions satisfying locality sensitive hashing schema. Moses proposed a triangle inequality that existing LSH families satisfy [17]. Since HeteSim could eventually be placed under the cosine similarity of hash vectors, we just need to prove $\theta(x_i, x_j) \leq \theta(x_i, y) + \theta(x_j, y)$: if x_i, x_j and y all lie in a plane, then it is obvious that the angle between x_i and x_j must be no greater than the sum of the angles between x_i and y and x_j and y; if x_i, x_j and y do not lie in a plane, and y' is the projection of y in the plane defined by x_i and x_j, then it holds that $\theta(x_i, x_j) \leq \theta(x_i, y') + \theta(x_j, y')$. Note the sum of angles between x_i and y and x_j and y are greater than those between x_i and y' and x_j and y', so we have $\theta(x_i, x_j) \leq \theta(x_i, y) + \theta(x_j, y)$. Therefore, HeteSim satisfies the locality sensitive hashing schema.

As the dataset scale of biological networks is always huge, and similarity computing is time-consuming, we use a relatively simple random hyperplane hash function as Definition 6 shows.

Definition 6. Hash Function

$$h_r(x) = \begin{cases} 1, & r \cdot x \geq 0 \\ 0, & r \cdot x < 0 \end{cases} \tag{4}$$

r is a d dimensional random vector with each value drawn from the standard Gaussian distribution $N(0, 1)$, if $r \cdot x \geq 0$, return 1, otherwise return 0. Then each d dimensional vector is hashed to one binary bit [16].

Given an integer m, we select m hash functions randomly and independently from the family defined in Definition 6, denoted as $H_m = \{h_{r_1}, \ldots, h_{r_m}\}$. By applying each of them to a dimensional vector x, we can map x to an m dimension vector in $\{0, 1\}^m$, denoted as $H_m(x)$. Then $H_m(x)$ is referred to the hash vector for x and m is the corresponding hash vector dimension. For any data set $D \in R^d, H_m(x)$ can generate a set consists of hash vectors, which is called the hash table of D. Given an integer t, we choose $H_m^1, H_m^2, \ldots H_m^t$ from H_m independently and randomly. Each hash family generates a $r_i (1 \leq i \leq t)$ dimensional hash table.

3.4 LSH-HeteSim

The biological networks dataset used in the prediction of interactions between drugs and targets is large-scale, usually consisted of numbers of biological databases. Moreover, mining the interactions between drugs and targets need large amount of similarity measure computations of high-dimensional vectors. These facts lead to that using the naive algorithms can be very time-consuming. Based on the characters above, we proposed an optimized algorithm based on LSH called LSH-HeteSim which can reduce a lot of similarity measure computation with the strategy that using a candidate subset generated by hashing reduces the computation times. As the relationship between the drug and the target is heterogeneous, the similarity measure we used in our method is HeteSim which is suitable for relevance search problem in heterogeneous networks as introduced in Sect. 3.2.

Algorithm 2. *LSH-HeteSim*

Input: data set D, m, t, mea-path P, query object p
Output: a set of object pairs with their similarity value

```
1.  Build LSH indexing for D as 3.3 introduced
2.    for each hash table T_{H_m^i} (1 ≤ i ≤ t) do
3.        Hashing query object p to a bucket B_i
4.        Add objects hashed to the target bucket B_i to set Q
5.    end for
6.    for each object q ∈ Q do
7.        HeteSimQuery(p, q)
8.        Add all (p, q) pairs and their similarity value into set R
9.    end for
10. Return R
```

The input data of Algorithm 2 is network dataset D, query object p, dimension of hash vector m, number of hash tables t and meta-path P, and the output result is a collection of objects which are in the candidate set with a similarity value. Firstly, Algorithm 2 creates a LSH indexing structure for given data set D (step 1); then it produces t hash tables cyclically, and eventually gets a collection of all objects which are mapped to the same hash bucket, as the candidate set Q (step 2–5); finally, it calculates the similarity between the query object p and each target object in candidate set Q using HeteSim, and returns a result set of all objects with a similarity value.

4 Experiments and Results

In order to completely inspect performance of our algorithm, we use several real biological networks as experiment data sets, rather than artificial data sets. Conducting experiments in real biological networks, it can direct compare capability and running time of our algorithm and existing methods.

4.1 Datasets

An integrated biological networks dataset Slap [6] is utilized as experimental material in this study. The network is constructed from 17 public data sources, which contains 305,792 nodes and 670,546 edges (Fig. 1). For the network, nodes are categorized into 10 types, in which 11 connections are existed. In addition, a single node is an instance of a corresponding type, for example: a node for drug Clofarabine (CID: 119182, Molecular Formula: C10H11ClFN5O3) is an instance of type Chemical Compound. A path is an instance of a corresponding meta-path, defined in Definition 2, for example: Troglitazone-Disease(776)-VEGE is an instance of the meta-path: Compound-Disease-Genes, here Troglitazone is a drug that can treat the disease 776 caused by VEGE.

4.2 Effectiveness

Assessing Drug Similarity. In Sect. 3.2, we have already mentioned when we choose the right meta-path which is to ensure the starting object type is the same as end object type, such as drugs, HeteSimQuery can be used to measure the similarity between objects with the same type. Here we use HeteSimQuery to cluster drugs. We took 40 kinds of drugs from 4 disease areas (headache, diabetes, HIV and asthma) to determine whether our method is able to distinguish drugs from different therapeutic areas. For each drug, we calculated its similarity value with the other 39 drugs. Then we selected all the drug-drug pairs whose similarity values were greater than a predefined threshold. In practice, drug-target interactions were visualized by the Cytoscape software [19], and functions of drug target genes were annotated through the iGepros server [21].

Since the path we used is CDGDC, namely two drugs can be regarded as similar when they can treat diseases caused by same genes. Therefore drugs related to same kind

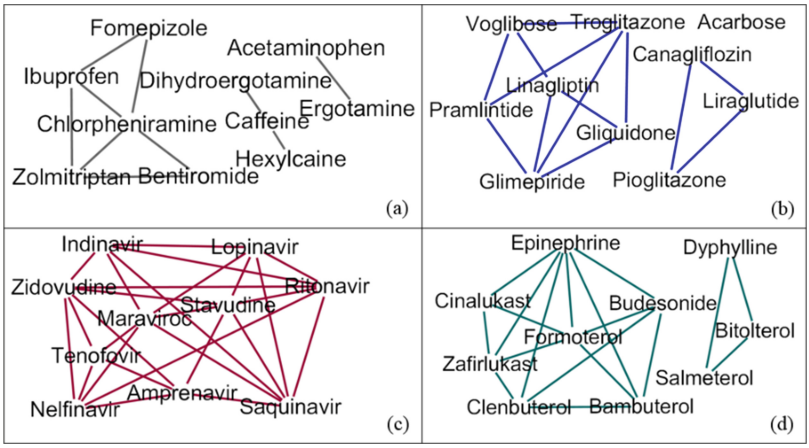

Fig. 3. Drug similarity network

of diseases have a higher probability to connect together. Our experiment results support this viewpoint. As shown in Fig. 3, the drugs in each group tend to have an ability to treat the same kind of disease, for example, the grey-colored drugs like Ibuprofen, Chlorpheniramine, Fomepizole can treat the headache and they connected together.

Comparison with SLAP. To further evaluate the drug-target interaction mining effect of the our optimized algorithm LSH-HeteSim, we compare LSH-HeteSim with the baseline algorithms HeteSimQuery and SLAP respectively. The experimental dataset used here is a subset extracted from the Slap dataset, including 1000 drugs, 127 targets and 3762 drug-target links. Then we have 127,000 drug target pair samples (3762 positive and 123,238 negative samples). The algorithm HeteSimQuery and SALP need to compute the similarities over the whole 127,000 drug-target pairs. However, LSH-HeteSim only needs to compute similarity for each query object with its corresponding candidate set. In our experiments, the parameter m and t are assigned 20 and 5 respectively, therefore LSH-HeteSim takes 22,170 times similarity computation in total.

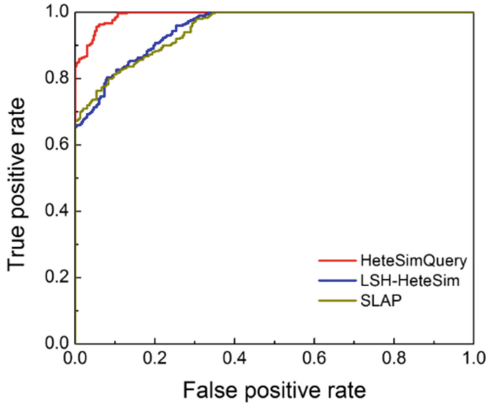

Fig. 4. ROC curves among different methods

To compare the accuracy effects of these three algorithms, we performed a ROC [20] (receiver operating characteristics) statistical analysis over the results. The ROC curves are shown in Fig. 4 which present achievable true positive rates (TP) with respect to all false positive rates (FP). The *AUC* (area under an ROC curve) values of HeteSimQuery, SLAP and LSH-HeteSim are 0.982, 0.940 and 0.943 respectively. Obviously, these three algorithms all have good prediction accuracy in the dataset. However, compared with our LSH-HeteSim algorithm, HeteSimQuery and SLAP algorithms require more running times. We will compare and describe in the experiment of Sect. 4.3. For drug target query in the large-scale biological data, time efficiency is very important. Our LSH-HeteSim algorithm reduced the similarity calculation times of high dimensional vector by LSH so that it can reduce the query time while ensure the prediction accuracy.

4.3 Efficiency

Efficiency is measured by the running time of algorithm. Since the hash functions are randomly picked, each experiment is repeated 10 times and the average is reported. The input of LSH-HeteSim algorithm has two parameters: the hash vector dimension m, the number of hash tables t. Here we take m and t into account to discuss how the running time changes.

As HeteSim is a path-dependent method, the running time is various when different path is selected. In out experiment, we use the meta-path: CGCDG (Compound-Gene-Compound-Disease-Gene). To better describe the running time with different parameters, we divide the experiments into two groups, and discuss the impact of parameters t and m on the running time.

Firstly, we randomly select a compound (noted as CID) as the query object such as CID = 5880, and the parameter m was assigned 20, then run programs for $t = 1, 2, 3, 4, 5$ respectively. Results of experiment are shown in Fig. 5(a). Clearly the trend of curve can be seen from the diagram, running time increases with the value of the parameter t added, mainly because a bigger value of t means more hash tables, and therefore more time to require for calculation.

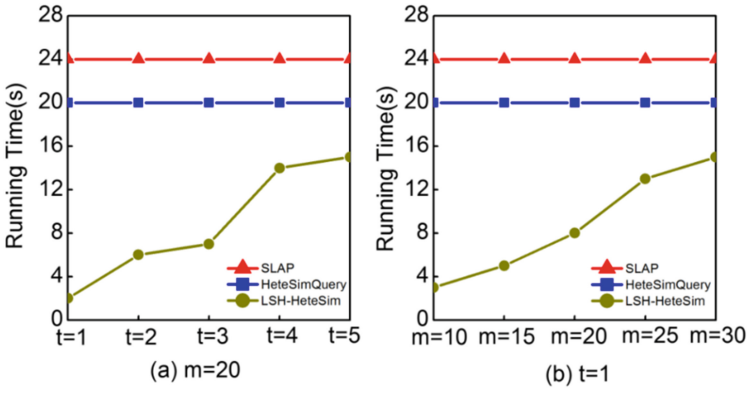

Fig. 5. Running time on Slap

Then we set parameter t as 1 and let the value of parameter m be 10, 15, 20, 25, and 30 respectively. Each experiment was repeated 10 times to take the average running time as the result. Obviously with the parameter m increases, the running time is growing, and this is mainly due to the increase in dimension of the random vector, it will take more time to calculate the similarity of two objects.

For a certain query object, the running times for SLAP and HeteSimQuery are almost fixed and more than LSH-HeteSim's, so their cures are linears. It is mainly due to our candidate sets generated by LSH, which greatly reduce the computation times of high-dimensional vectors. Considering results of effectiveness and efficiency assesses, the LSH-HeteSim algorithm has less running time compared with the state-of-art methods, and its prediction accuracy is still comparable to these methods. In addition, there is no fixed range of the two parameters, and the suitable values of the two parameters should be assigned through value trials. With the increase of the parameters m and t, the times of similarity calculations and the dimension of feature vectors increase. Despite the prediction accuracy will rise, however, the running time is also rapidly increasing.

5 Discussion and Conclusion

For the LSH-HeteSim algorithm, it accelerates computing of similarity search in high-dimensional space through the LSH method, which results in a slightly decrease of search accuracy (experiments in Fig. 4). In the future, we will use the MP-LSH method [22] instead of the LSH method to optimize our algorithm. In this way, accuracy of the LSH-HeteSim algorithm can be improved, and its less running time characteristics can be kept. In addition, the meta-path CGCDG used in this study is based on certain biological information, while meta-paths coming from other biological information should be inspected in the future.

In this study, we proposed an efficient drug-target interaction mining algorithm for heterogeneous biological networks called LSH-HeteSim. Experiment results show that our proposed algorithm can effectively predict interactions between drugs and targets. Specially, for larger-scale biological data, LSH-HeteSim has less running time compared with the state-of art methods.

Acknowledgment. The work was supported in part by the National Natural Science Foundation Project of China under Grant. No. 61170096 and Research Program of Shanghai Science and Technology Committee under Grant. No. 12511502403. The authors gratefully acknowledge the support of SA-SIBS Scholarship Program.

References

1. Hanzlik, R.P., Koen, Y.M., Theertham, B., et al.: The reactive metabolite target protein database (TPDB)—a web-accessible resource. BMC Bioinf. **8**(1), 95 (2007)
2. Chan, S.Y., Loscalzo, J.: The emerging paradigm of network medicine in the study of human disease. Circul. Res. **111**(3), 359–374 (2012)

3. Chen, L., Lu, J., Luo, X., et al.: Prediction of drug target groups based on chemical-chemical similarities and chemical-chemical/protein connections. Biochim. Biophys. Acta (BBA)-Proteins and Proteomics **1844**, 207–213 (2013)
4. Campillos, M., Kuhn, M., Claude, G., et al.: Drug target identification using side-effect similarity. Science **321**(5886), 263–266 (2008)
5. He, Z., Zhang, J., Shi, X.H., et al.: Predicting drug-target interaction networks based on functional groups and biological features. PloS one **5**(3), e9603 (2010)
6. Chen, B., Ying, D., David, J.W.: Assessing drug target association using semantic linked data. PLoS Comput. Biol. **8**(7), e1002574 (2012)
7. Yamanishi, Y., Kotera, M., Kanehisa, M., et al.: Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics **26**(12), 246–254 (2010)
8. Fakhraei, S., Louiqa, L., Lise, G.: Drug-target interaction prediction for drug repurposing with probabilistic similarity logic. In: Proceedings of the 12th International Workshop on Data Mining in Bioinformatics. ACM (2013)
9. Sun, Y.Z., Han, J.W., Yan, X.F., et al.: PathSim: meta path-based top-k similarity search in heterogeneous information networks. In: VLDB'11 (2011)
10. Shi, C., Kong, X.N., Yu, P.S., et al.: Relevance search in heterogeneous networks. In: Proceedings of the 15th International Conference on Extending Database Technology. ACM (2012)
11. Palma, G., Viadl, M.-E., Haag, L., et al.: Measuring relatedness between scientific entities in annotation datasets. In: Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics. ACM (2013)
12. Getoor, L., Diehl, C.P.: Link mining: a survey. ACM SIGKDD Explor. Newslett. **7**(2), 3–12 (2005)
13. Yu, H.Y., Braun, P., Yildirim, M.A., et al.: High-quality binary protein interaction map of the yeast interactome network. Science **322**(5898), 104–110 (2008)
14. Datar, M., Immorlica, N., Indyk, P., et al.: Locality-sensitive hashing scheme based on p-stable distributions. In: Proceedings of the Twentieth Annual Symposium on Computational Geometry. ACM (2004)
15. Jegou, H., Matthijs, D., Cordelia, S.: Product quantization for nearest neighbor search. IEEE Trans. Pattern Anal. Mach. Intell. **33**(1), 117–128 (2011)
16. Kishore, S.: Accelerated clustering through locality-sensitive hashing. Diss. Massachusetts Institute of Technology (2012)
17. Charikar, M.S.: Similarity estimation techniques from rounding algorithms. In: Proceedings of the thirty-fourth annual ACM symposium on Theory of computing. ACM (2002)
18. SLAP for Drug Target Prediction. http://cheminfov.informatics.indiana.edu:8080/slap
19. Saito, R., Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., Lotia, S., Pico, A.R., Bader, G.D., Ideker, T.: A travel guide to Cytoscape plugins. Nat. Methods **9**(11), 1069–1076 (2012)
20. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2005)
21. Zheng, G., Wang, H., Wei, C., Li, Y.: iGepros: an integrated gene and protein annotation sever for biological nature exploration. BMC Bioinf. **12**(Suppl 14), S6 (2011)
22. Lv, Q., Josephson, W., Wang, Z., et al.: Multi-probe LSH: efficient indexing for high-dimensional similarity search. VLDB Endowment, pp. 950–961 (2007)