Chapter 19

Tripartite Network-Based Repurposing Method Using Deep Learning to Compute Similarities for Drug-Target Prediction

Nansu Zong, Rachael Sze Nga Wong, and Victoria Ngo

Abstract

The drug discovery process is conventionally regarded as resource intensive and complex. Therefore, research effort has been put into a process called drug repositioning with the use of computational methods. Similarity-based methods are common in predicting drug-target association or the interaction between drugs and targets based on various features the drugs and targets have. Heterogeneous network topology involving many biomedical entities interactions has yet to be used in drug-target association. Deep learning can disclose features of vertices in a large network, which can be incorporated with heterogeneous network topology in order to assist similarity-based solutions to provide more flexibility for drug-target prediction. Here we describe a similarity-based drug-target prediction method that utilizes a topology-based similarity measure and two inference methods based on the similarities. We used DeepWalk, a deep learning method, to calculate the vertex similarities based on Linked Tripartite Network (LTN), which is a heterogeneous network created from different biomedical-linked datasets. The similarities are further used to feed to the inference methods, drug-based similarity inference (DBSI) and target-based similarity inference (TBSI), to obtain the predicted drug-target associations. Our previous experiments have shown that by utilizing deep learning and heterogeneous network topology, the proposed method can provide more promising results than current topology-based similarity computation methods.

Key words Drug-target association, Tripartite network, Deep learning, DeepWalk, Similarity-based drug-target prediction, Heterogeneous network topology, Bipartite network

1 Introduction

The experimental process of drug discovery (in vitro) is very time consuming, complex, and expensive. Additionally, the success rate of empirically locating the associations between drugs and targets is decreasing. Therefore, the pharmaceutical industry has shifted toward novel computational methods as a new strategy toward drug repurposing. One of the most important parts of drug repurposing is identifying and verifying the interactions between drugs and targets. Current efforts encompass identifying drug-target associations and developing chemical compounds that utilize "druggable" proteins [1]. Drugs specifically bind targets to modify

Quentin Vanhaelen (ed.), *Computational Methods for Drug Repurposing*, Methods in Molecular Biology, vol. 1903,
https://doi.org/10.1007/978-1-4939-8955-3_19, © Springer Science+Business Media, LLC, part of Springer Nature 2019

their biochemical and/or biophysical behavior and lead to desirable chemical reactions. Hence, researchers pay attention to a number of completed pharmacological profiles of certain desired target proteins which leave some small molecules to be rarely studied [2]. The diverse associations between drugs and targets create a highly interconnected cellular network. In the past years, the pharmaceutical industry followed the "one molecule-one target-one disease" standard, which explained that specific drugs would act on one target for a specific disease [3]. Since complex diseases may require addressing multiple targets, however, this standard has been challenged and the industry is exploring poly-pharmacology, where the design focuses on multiple targets instead of one individual target, and repurposing existing drugs, such as anticancer drugs imatinib (Gleevec) [1, 4]. There are still many limitations to the understanding of drug-target associations compared to the number of chemical compounds and proteins already discovered. This gap encourages the study of the predictions of drug-target associations between existing drugs and their targets [5, 6].

Computational methods are reliable approaches that would allow researchers to devote less time on experiments and give them estimates of success before experiment initiation. Previously, computational predictions using docking simulation [7] and text mining methods [8] were not scalable and sufficient enough to analyze proteins that did not contain three-dimensional structure information. Additionally, the scientific literature databases that contained protein and gene names were too complex with disorganized information, which posed a challenge on text mining. As a solution to these difficulties, researchers utilized diverse machine learning methods to predict drug-target associations. To further enhance the results received, similarity measures were key to the success of these methodologies. For example, in order to compute the weighting of potential associations, the similarity measures of drug-drug and target-target pairs were used [4, 6]. Analyzing the association between two components offered flexible solutions to practical scenarios and yielded to the best combinations [9, 10]. Generally, similarity measure utilized information from genome sequences [6, 11, 12], pharmacological features [10], and chemical structures [6].

Studies have shown that in heterogeneous networks, information on topological interactions between biomedical entities can be beneficial for the prediction process [4, 13–16]. Yet, topology-based methods cannot be used in the existing similarity-based methods because they cannot compute topological similarities for biological entities. Therefore, deep learning is needed to extract features of vertices in a large network. It can also be used to generate topological similarities of two vertices [17, 18]. With deep learning, the method of drug-target prediction is significantly improved as currently existing similarity-based methods can be reused and integrated.

DeepWalk, a similarity-based drug-target prediction with deep learning algorithm implemented in this study, uses the topology of a heterogeneous network called Tripartite Linked Network (TLN) to calculate the similarities of drug-drug and target-target pairs [19]. The "guilty-by-association" principle is used to compare the resulting similarity measure with drug-target association by using drug-drug and target-target similarities as the input [20]. This method is expected to compute promising results in the drug-target association prediction, for example, a 98.96% AUC ROC score with a tenfold cross-validation and a 99.25% AUC ROC score with Monte Carlo cross-validation can be achieved [21].

2 Materials

We constructed a tripartite network that included three types of vertices: drugs, targets, and diseases. Correspondingly, three types of associations, drug-target, drug-disease, and disease-target associations, were used as the edges to connect the vertices. The network is constructed based on the knowledge (i.e., associations) from two existed knowledge base, DrugBank [22], and human disease network [23]. In practice, we used the linked data version of the two databases. The linked data version of DrugBank uses the version 3 of the original database generated in 2011; we downloaded the data (http://wifo5-03.informatik.uni-mannheim.de/drugbank/) and extracted 4553 targets, 4408 drugs, and 12,045 drug-target associations from the database. Linked data version of Diseasome was downloaded from (http://wifo5-03.informatik.uni-mannheim.de/diseasome/), and we extracted 1452 diseases and 8201 drug-disease associations from the data. To establish the disease-target association for the network, we mapped the genes in DrugBank and Diseasome based on four databases, Bio2RDF [24], UniProt [25], HGNC [26], and OMIM [27]. The entities with the same knowledge base ID were considered as the same entity and mapped with "owl:*sameAS*" (*see* **Note 1**). Since linked datasets were used to build the network, we called our network Linked Tripartite Network (LTN) and demonstrated the network in Fig. 1.

In the network, three kinds of vertices are represented with three colors and shapes, which are *ellipse* and *green* for drugs, *rectangle* and *blue* for targets, and *hexagon* and *yellow* for diseases (*see* **Note 2**). There are 395 connected components in total, in which the largest one contains 9283 vertices, each of them having an average of 4.5 neighbors. We have demonstrated the following results of network analysis produced by Cytoscape [28], which are (a) node degree distribution, (b) average clustering coefficient

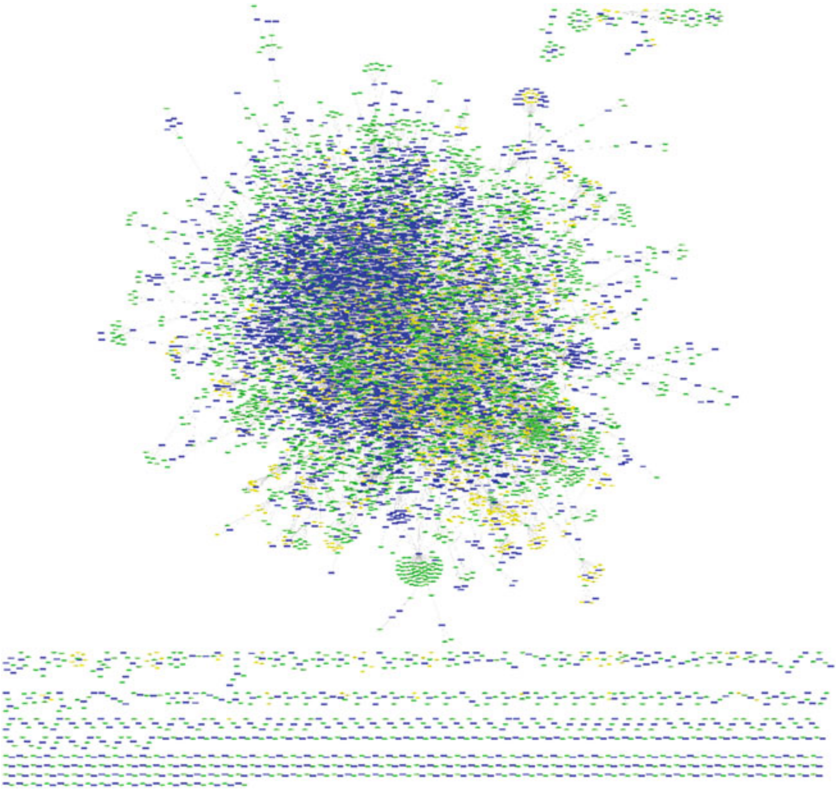

Fig. 1 Visualization of Linked Tripartite Network (LTN)

distribution, (c) topological coefficient, (d) neighborhood connectivity distribution, (e) betweenness centrality, and (f) closeness centrality in Figs 2, 3, and 4.

3 Method

We separated our drug-target discovery strategy into two parts: (1) association discovery and (2) similarity computation. Association discovery that is conducted on-the-fly takes a drug or a target as the input and returns a list of drug-target associations with the probability scores. The probability scores are computed based on two popular rule-based inference methods, drug-based similarity inference (DBSI) and target-based similarity inference (TBSI) [4, 6]. The two methods induce the possible potential drug-target associations based on "guilt-by-association" principle [20] that postulates a potential association can be established between *node A* and *node 1* if *node A* is similar to *node B* and there is an existing association between *node B* and *node 1* (*see* **Note 3**).

Given a pair of a drug d_i and a target t_j, DBSI gives a probability score for such association as

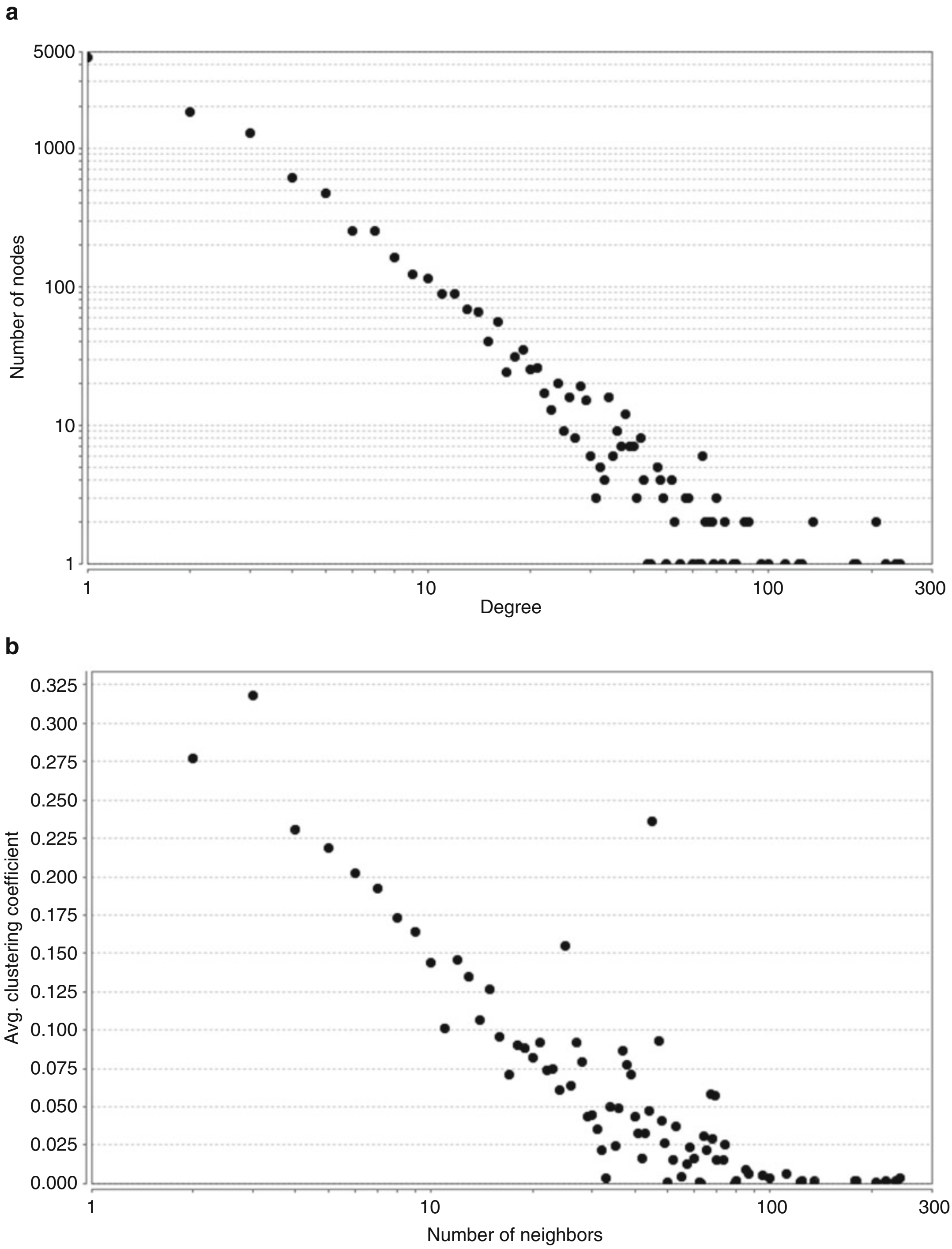

Fig. 2 Network analysis 1 of LTN. (**a**) Node degree distribution. (**b**) Average clustering coefficient distribution

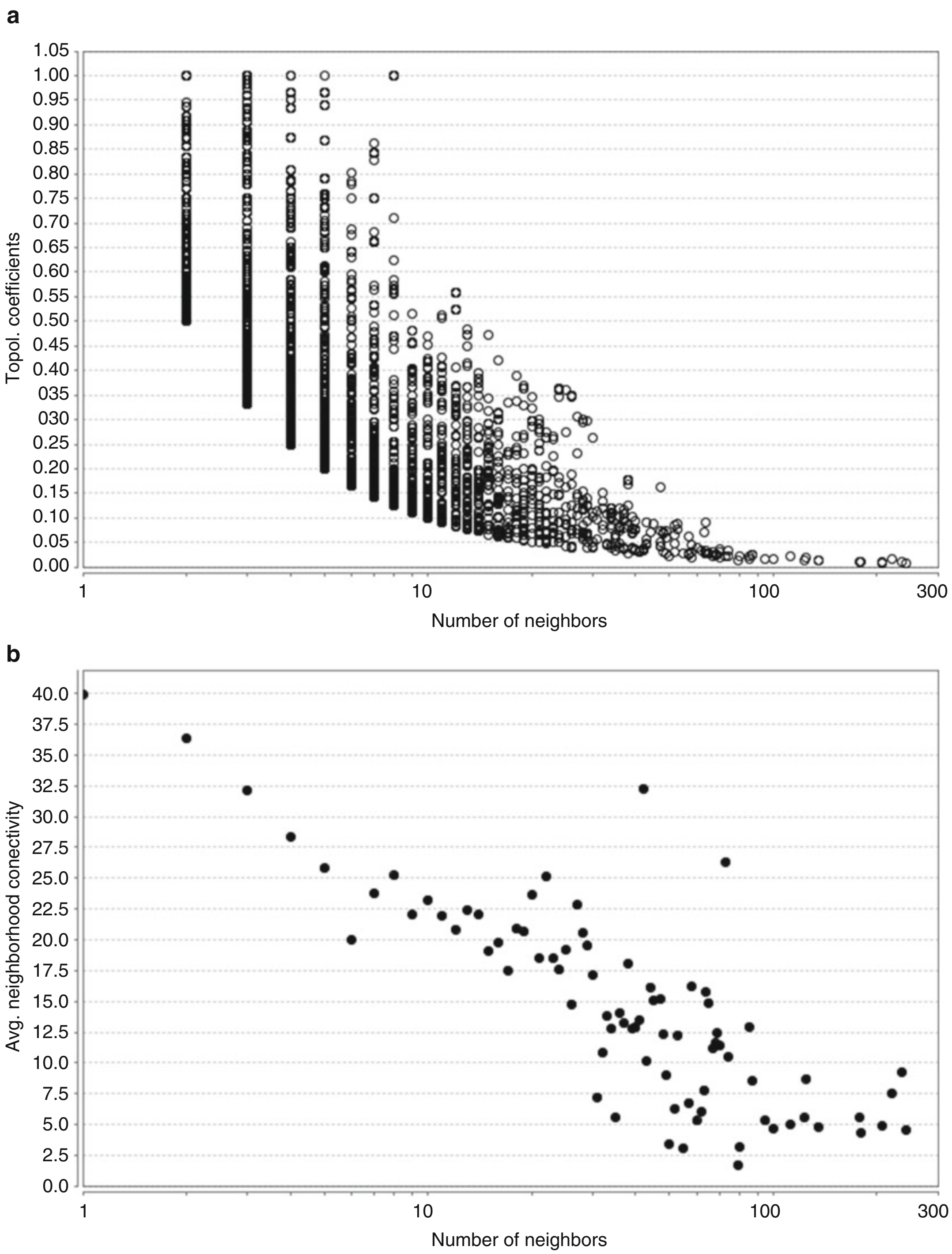

Fig. 3 Network analysis 2 of LTN. (**a**) Topological coefficient. (**b**) Neighborhood connectivity distribution

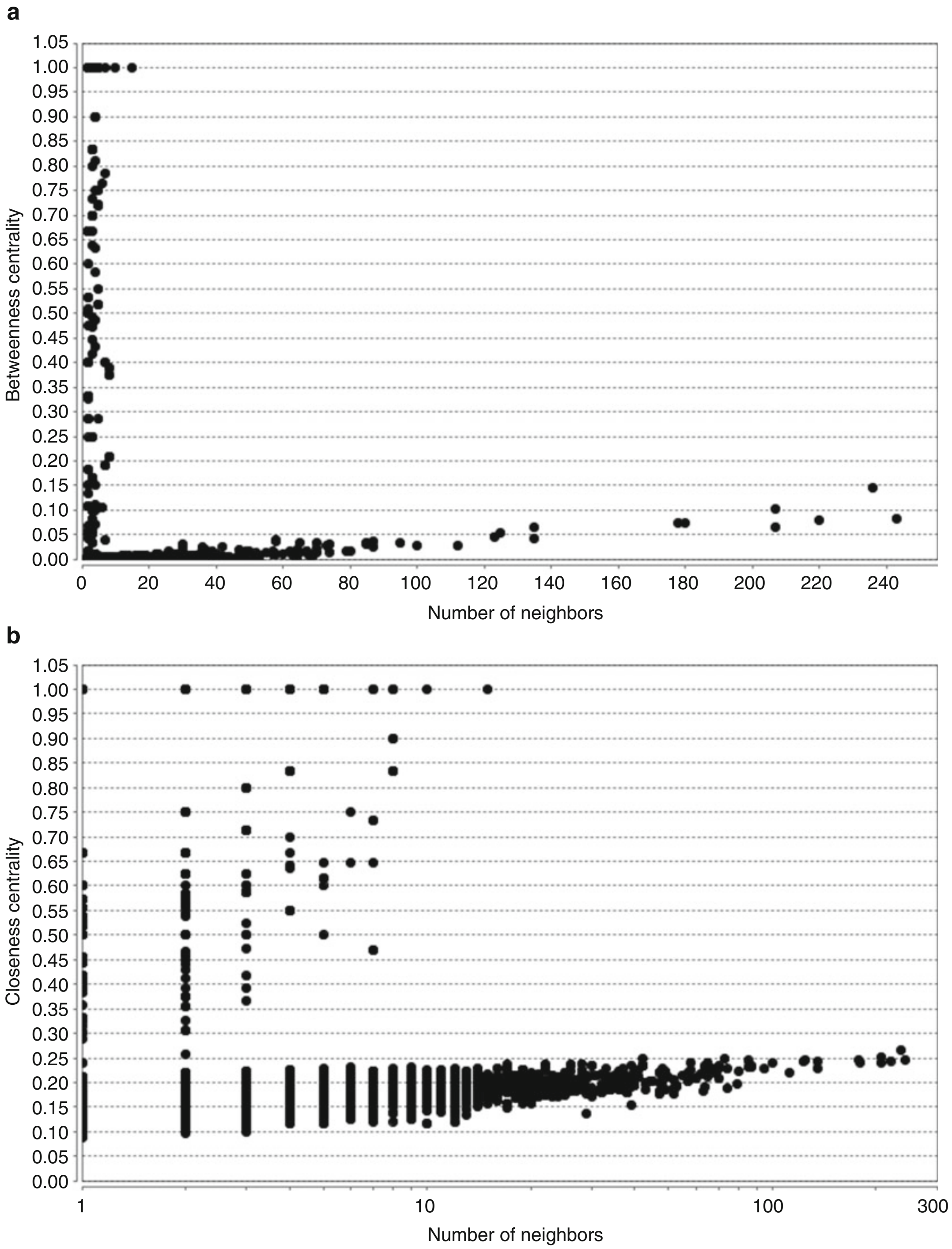

Fig. 4 Network analysis 3 of LTN. (**a**) Betweenness centrality. (**b**) Closeness centrality

$$P\left(d_i, t_j\right)_{\mathrm{DBSI}} = \frac{\sum_{l=1, l\neq i}^{n} \mathrm{sim}\left(d_i, d_l\right) a_{l,j}}{\sum_{l=1, l\neq i}^{n} \mathrm{sim}\left(d_i, d_l\right)} \tag{1}$$

where sim(d_i, d_l) is the similarity between d_i and d_l, and $a_{l,j} = 1$ if there is an existing association between d_l and t_j; otherwise $a_{l,j} = 0$.

Similarly, TBSI gives a probability score for such association as

$$P\left(t_j, d_i\right)_{\mathrm{TBSI}} = \frac{\sum_{l=1, l\neq j}^{m} \mathrm{sim}\left(t_j, t_l\right) a_{i,l}}{\sum_{l=1, l\neq j}^{m} \mathrm{sim}\left(t_j, t_l\right)} \tag{2}$$

where sim(t_j, t_l) is the similarity between t_j and t_l, and $a_{i,l} = 1$ if there is an existing association between d_i and t_l; otherwise $a_{i,l} = 0$ (*see* **Note 4**).

To compute the similarity used for the above two equations, the vertices are represented as d dimensional vectors, and the similarity between two vertices is computed based on the cosine similarity as follows:

$$\mathrm{sim}(u, v) = \frac{\sum_{k=1}^{d} u_k v_k}{\sqrt{\sum_{k=1}^{d} u_k^2}\sqrt{\sum_{k=1}^{d} v_k^2}} \tag{3}$$

where d is the dimension and u_i and v_i are the components of vector u and v, respectively.

The node vector index is computed by a deep learning method called DeepWalk [18], which takes the network structure to vertices for computing the similarity between two vertices. DeepWalk uses truncated random walks to get latent topological information of the network and obtains the vector representation of the vertices by maximizing the probability of a next vertex given the previous vertices in these walks. We demonstrate the implementation of DeepWalk in pseudocode 1 (*see* Table 1). To compute the DeepWalk score of the vertices, window size w, vector size d, walks per vertex γ, walk length t, and a network $G(V, E)$ are needed as the input, and a list of vectors Φ are the results for the corresponding vertices in the network (*see* **Note 5**). Firstly, for each walk, each vertex u_i in V will conduct a t length of random walk based on the edges in the network $G(V, E)$. As a result, each walk will generate a walk path ω_{u_i} that includes all the steps (lines 3–5). With ω_{u_i} and window size w, a skip-gram model is used to update the vector Φ_{u_i}. More especially, for each vertex u_j belonging to walk ω_{u_i}, and each vertex u_k that is belonging to walk ω_{u_i} as well as within the range of w to u_j, a cost function $\mathrm{J}(\Phi_{u_i})$ will be used to update Φ_{u_i} (lines 6–8) with the learning rate α as

$$\Phi_{u_i} = \Phi_{u_i} - \alpha \frac{\partial \mathrm{J}(\Phi_{u_i})}{\partial \Phi_{u_i}} \tag{4}$$

where $\mathrm{J}(\Phi_{u_i}) = -\log Pr\left(u_k | \Phi_{u_j}\right)$. Based on hierarchical softmax, $Pr\left(u_k | \Phi_{u_j}\right)$ can be approximated as

Table 1
Pseudocode 1: DeepWalk computation

```
Input: window size $w$, vector size $d$, walks per vertex $\gamma$, walk length $t$, learning rate $\alpha$, a network $G(V, E)$
Output: Matrix $\Phi$

Initialization $\Phi$.
Generate binary tree $T$ from V.
For walk $i$ in $\gamma$ do.
    For vertex $u_i$ in V do.
    $\omega_{u_i}$:=Random Walk (G, $u_i$, t)
    For vertex $u_j$ in $\omega_{u_i}$ do.
    For vertex $u_k$ in window $\omega_{u_i}[j - w, j + w]$ do.
    update ($\Phi_{u_i}$),
```

$$Pr\left(u_k|\Phi_{u_j}\right) = \Pi_{l=1}^{log|V|} Pr\left(b_l|\Phi_{u_j}\right) \tag{5}$$

where b_l is one of the tree nodes in the path to the node u_j that modeled in a binary tree index generated from the network $Pr\left(b_l|\Phi_{u_j}\right)$ is computed as

$$Pr\left(b_l|\Phi_{u_j}\right) = {}^{1}\Big/\left(1+e^{-\Phi_{u_j}\cdot\Psi_{b_l}}\right) \tag{6}$$

where Ψ_{b_l} is the vector representation of the parent node of tree node b_l.

With the node vectors computed with DeepWalk, we can conduct the following method demonstrated in pseudocode 2 (*see* Table 2) to return a list of drug-target association on-the-fly for drug-target association prediction (*see* **Note 6**). We first need to extract drug-target associations, a list of drugs and a list of targets from LTN (lines 1–3). If the input query is a drug, each drug d_i in the drug list will be computed with the vector cosine similarity based on the node vector index, and the similarity will be linear combined to the existed probability score of the drug-target pair (Q, t_j) if the target t_j has an association with d_i (lines 4–9). Similarly, if the input query is a target, each target t_i in the target list will be computed with the vector cosine similarity based on the node vector index, and the similarity will be linear combined to the existed probability score of the drug-target pair (d_j, Q) if the drug d_j has an association with t_i (lines 10–15) (*see* **Note** 7).

4 Notes

1. To map the entities in different databases, the common third-party IDs are used for mapping in our work, such as Bio2RDF [24], UniProt [25], HGNC [26], and OMIM [27]. However,

Table 2
Pseudocode 2: drug-target association prediction

```
Input: Linked Tripartite network G, Node vector Index I, Query Q (drug or target)
Output: a list of drug-target associations with probability scores M

Existed drug-target associations $A$:= extracted from $G$.
Existed drugs $D$: = extracted from $G$.
Existed targets $T$: = extracted from $G$.
If $Q$ is a drug then.
    For $d_i$ in $D$ do.
        sim $(d_i, Q)$ = cosine _ similarity (vec($d_i$), vec($Q$))
        , where vec($d_i$) ≔ extracted from $I$,vec($Q$) ≔ extracted from $I$
        For $t_j$ associated with $d_i$ do.
            $M(Q, t_j) + = \text{sim}(d_i, Q)$
Else If $Q$ is a target then.
    For $t_i$ in $T$ do.
        sim $(Q, t_i)$ = cosine _ similarity(vec($Q$), vec($t_i$))
        , where vec($t_i$) ≔ extracted from $I$,vec($Q$) ≔ extracted from $I$
        For $d_j$ associated with $t_i$ do.
            $M(d_j, Q) + = \text{sim}(Q, t_i)$

Return M as the prediction results.
```

other common IDs might also be used for mapping in the future. To preserve the precision in mapping, we only use the common ID in one step while ignoring the N-step transitive common ID, which can be considered to further reduce the repentant vertices in the networks. Using other mapping methods by computing the label or description similarity between two entities can be considered as well [29].

2. To compute the entity similarity based on topology of the network, vectorization should be conducted to obtain the vector representation of the vertices (i.e., entities). Therefore, all the entities should be represented in a same data space, which requires the unique vertex for each entity. All the different entities originate from the different databases but represent the same concept which should be mapped to a designated entity in a specific dataspace. In our work, we used the drug entities and target entities from DrugBank as our dataspace for drugs and targets and used disease entities from Diseasome for diseases.

3. The proposed method can only be used to predict the potential associations between the drugs or targets that have drug-target associations. In another words, to predict a target with a given a drug as an input, the prediction can fail based on DBSI if the target does not have any target-drug association existed in the network. Similarly, to predict a drug with a given target as an

input, the prediction can fail based on TBSI if the drug does not have any target-drug association existed in the network.

4. The confidence (i.e., likelihood) of the drug-target prediction is given by a normalized value from the cosine similarity as

$$\text{Nomarlized confidence}_{\text{DBSI}}\left(d_i, t_j\right) = \frac{\text{Confidence}_{\text{DBSI}}\left(d_i, t_j\right) - \text{Max}\left(d_i, \cdot\right)}{\text{Max}\left(d_i, \cdot\right) - \text{Min}\left(d_i, \cdot\right)} \tag{7}$$

$$\text{Nomarlized confidence}_{\text{DBSI}}\left(d_i, t_j\right) = \frac{\text{Confidence}_{\text{TBSI}}\left(d_i, t_j\right) - \text{Max}\left(\cdot, t_j\right)}{\text{Max}\left(\cdot, t_j\right) - \text{Min}\left(\cdot, t_j\right)} \tag{8}$$

The two equations give a confidence score between 0 and 1.

5. We computed DeepWalk with deeplearning4j library (http://deeplearning4j.org/), which is a deep learning open source for JAVA. Other tools for DeepWalk can be found for C++ (https://github.com/xgfs/deepwalk-c) and Python (https://github.com/phanein/deepwalk).

6. Similar to the classification method in machine learning, to give a more straightforward result of the prediction, a threshold can be used in the application to simply give a binary result for a prediction.

7. The method introduced is component based. The two components introduced can be replaced with other similar methods. For example, other similarity measure for computing the similarity of the chemical structure of the drugs and the similarity of the genomic sequence of the targets can be used to replace the original similarity computation module. The inferences component can also be replaced with some classification algorithms in machine learning.

References

1. Yıldırım MA, Goh K-I, Cusick ME, Barabási A-L, Vidal M (2007) Drug—target network. Nat Biotechnol 25(10):1119–1126
2. Vogt I, Mestres J (2010) Drug-target networks. Mol Inform 29(1-2):10–14
3. Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 4(11):682
4. Cheng F, Liu C, Jiang J, Lu W, Li W, Liu G, Zhou W, Huang J, Tang Y (2012) Prediction of drug-target interactions and drug repositioning via network-based inference. PLoS Comput Biol 8(5):e1002503
5. Ding H, Takigawa I, Mamitsuka H, Zhu S (2014) Similarity-based machine learning methods for predicting drug–target interactions: a brief review. Brief Bioinform 15(5):734–747
6. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M (2008) Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 24(13):i232–i240
7. Cheng AC, Coleman RG, Smyth KT, Cao Q, Soulard P, Caffrey DR, Salzberg AC, Huang ES (2007) Structure-based maximal affinity

model predicts small-molecule druggability. Nat Biotechnol 25(1):71–75

8. Zhu S, Okuno Y, Tsujimoto G, Mamitsuka H (2005) A probabilistic model for mining implicit 'chemical compound–gene'relations from literature. Bioinformatics 21(suppl 2): ii245–ii251
9. Perlman L, Gottlieb A, Atias N, Ruppin E, Sharan R (2011) Combining drug and gene similarity measures for drug-target elucidation. J Comput Biol 18(2):133–145
10. Yamanishi Y, Kotera M, Kanehisa M, Goto S (2010) Drug-target interaction prediction from chemical, genomic and pharmacological data in an integrated framework. Bioinformatics 26(12):i246–i254
11. Bleakley K, Yamanishi Y (2009) Supervised prediction of drug–target interactions using bipartite local models. Bioinformatics 25 (18):2397–2403
12. Jacob L, Vert J-P (2008) Protein-ligand interaction prediction: an improved chemogenomics approach. Bioinformatics 24 (19):2149–2156
13. Palma G, Vidal M-E, Raschid L (2014) Drug-target interaction prediction using semantic similarity and edge partitioning. In: International semantic web conference. Springer, pp 131–146
14. Wang W, Yang S, Li J (2013) Drug target predictions based on heterogeneous graph inference. In: Pacific symposium on biocomputing. Pacific symposium on biocomputing. NIH Public Access, p 53
15. Chen X, Liu M-X, Yan G-Y (2012) Drug–target interaction prediction by random walk on the heterogeneous network. Mol BioSyst 8 (7):1970–1978
16. Chen B, Ding Y, Wild DJ (2012) Assessing drug target association using semantic linked data. PLoS Comput Biol 8(7):e1002574
17. Tang J, Qu M, Wang M, Zhang M, Yan J, Line MQ (2015) Large-scale information network embedding. In: Proceedings of the 24th international conference on world wide web. ACM, pp 1067–1077
18. Perozzi B, Al-Rfou R, Deepwalk SS (2014) Online learning of social representations. In: Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp 701–710
19. Bizer C, Heath T, Berners-Lee T (2009) Linked data-the story so far. In: Semantic services, interoperability and web applications: emerging concepts, pp 205–227
20. Bass JIF, Diallo A, Nelson J, Soto JM, Myers CL, Walhout AJ (2013) Using networks to measure similarity between genes: association index selection. Nat Methods 10(12):1169–1176
21. Zong N, Kim H, Ngo V, Harismendy O (2017) Deep mining heterogeneous networks of biomedical linked data to predict novel drug–target associations. Bioinformatics 33 (15):2337–2344
22. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, Tzur D, Gautam B, Hassanali M (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36(suppl 1):D901–D906
23. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L (2007) The human disease network. Proc Natl Acad Sci 104 (21):8685–8690
24. Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J (2008) Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform 41(5):706–716
25. Consortium U (2008) The universal protein resource (UniProt). Nucleic Acids Res 36 (suppl 1):D190–D195
26. Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H (2001) The HUGO gene nomenclature committee (HGNC). Hum Genet 109(6):678–680
27. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA (2005) Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33(suppl 1):D514–D517
28. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504
29. Volz J, Bizer C, Gaedke M, Kobilarov G (2009) Silk-a link discovery framework for the web of data. LDOW 538