Peng *et al. BMC Systems Biology* 2018, **12**(Suppl 2):18
https://doi.org/10.1186/s12918-018-0539-0

BMC Systems Biology

RESEARCH Open Access

Improving the measurement of semantic similarity by combining gene ontology and co-functional network: a random walk based approach

Jiajie Peng[1,2,3*], Xuanshuo Zhang[1], Weiwei Hui[1], Junya Lu[1], Qianqian Li[1], Shuhui Liu[1] and Xuequn Shang[1,2]

From The 28th International Conference on Genome Informatics
Seoul, Korea. 31 October - 3 November 2017

Abstract

Background: Gene Ontology (GO) is one of the most popular bioinformatics resources. In the past decade, Gene Ontology-based gene semantic similarity has been effectively used to model gene-to-gene interactions in multiple research areas. However, most existing semantic similarity approaches rely only on GO annotations and structure, or incorporate only local interactions in the co-functional network. This may lead to inaccurate GO-based similarity resulting from the incomplete GO topology structure and gene annotations.

Results: We present NETSIM2, a new network-based method that allows researchers to measure GO-based gene functional similarities by considering the global structure of the co-functional network with a random walk with restart (RWR)-based method, and by selecting the significant term pairs to decrease the noise information. Based on the EC number (Enzyme Commission)-based groups of yeast and Arabidopsis, evaluation test shows that NETSIM2 can enhance the accuracy of Gene Ontology-based gene functional similarity.

Conclusions: Using NETSIM2 as an example, we found that the accuracy of semantic similarities can be significantly improved after effectively incorporating the global gene-to-gene interactions in the co-functional network, especially on the species that gene annotations in GO are far from complete.

Keywords: Gene Ontology; Semantic similarity; Random walk with restart

Background

Recently, significant improvement in high-throughput biology technologies has led to an exponential increase in biological data. Gene Ontology (GO) is one of the most popular bioinformatics resources used to interpret the result of biological experiment. GO provides structured, controlled vocabulary of terms to describe genes by three types of attributes that are molecular function, biological process and cellular component [1]. In each category, terms are structured as a directed acyclic graph (DAG). GO provides a convenient and important way to study functional similarity. GO-based semantic similarity has been successfully used in many research areas, such as gene function prediction [2–5], gene network analysis [6, 7], homology analysis [8], gene association visualization [9] and missing value imputation [10, 11].

In the past decade, a lot of approaches have been proposed to calculate gene functional similarity based on gene ontology [12–23]. Based on the information used in similarity calculation, these measurements can be loosely classified into four groups: path length-based methods, node-based methods, integrative methods and network-based methods.

*Correspondence: jiajiepeng@nwpu.edu.cn
[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China
[2]Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University, Ministry of Industry and Information Technology, Xi'an, China
Full list of author information is available at the end of the article

The methods in the edge-based group calculate similarity by considering the topology structure information of GO [24, 25]. A recently proposed approach, named Relative Specificity Similarity (RSS), takes two types of length information into account: the edge length from given term pair to their closest leaf terms; and the edge length to their lowest common ancestor (LCA) [25]. The experiment result shows that this method is superior in correlation with sequence and Pfam similarities. However, the edge-based methods are fully relied on the topology of GO DAG. This type of methods cannot differentiate the terms at the same topological level [14].

For the node-based methods, the approaches rely on the specific taxonomy. One of the proposed approaches exploit the information content (IC) of the most informative common ancestor (MICA) to measure the similarity between two GO terms [26]. Let t be a MICA term. We calculated its IC as $-log(|G_t|/|G_{root}|)$. G_t and G_{root} represent gene sets annotated to term t and *root* respectively. This method is further improved by taking the path length from the term pair to its MICA into account [12]. The evaluation test shows that the results are consistent with protein sequence similarities. However, node-based approaches only take the annotations into account, ignoring the topology information of the GO.

In the integrative group, the approaches are proposed to use more information in GO. Hybrid Relative Specificity Similarity (HRSS) uses four types of information (information content, structure topology, annotations and MICA) to calculate the semantic similarity [25]. InteGO method proposed a rank-based method to integrate multiple existing similarity methods, called seed methods, to consider more aspects of GO [17]. InteGO2 method selects the most appropriate methods from a set of methods by a voting method and integrates these selected methods based on a metaheuristic search method [9]. The evaluation test shows that the integrative method performs better than the seed method. However, all these methods are only based on the GO, neglecting the inaccurate representation and missing information of GO. For example, 37% of the Arabidopsis genes have experimental annotations of all three domains of GO [27]. Therefore, low-quality similarity may result from the incomplete information in GO.

A network-based method, called NETSIM, was recently proposed to address these problems by integrating gene-gene associations and GO topology structure and annotations [19]. The experiment based on metabolic reaction map shows that semantic similarity can be enhanced by incorporating gene-gene associations. Unfortunately, only part of the information in gene co-function network was used, since NETSIM only considered the direct link in the network. Other than the directly connected gene pairs, the indirect gene-gene interactions contained in the gene co-function network should also be considered. However, considering indirect interactions may also import the noise information.

In this paper, we proposed a novel network-based method named NETSIM2, by considering both direct and indirect interactions in the gene co-function network with a random walk based method, and by selecting the significant term pairs for similarity calculation to decrease the effect of the imported noise information. Comparing with the existing approaches, NETSIM2 has the following advantages:

- Comparing with the state-of-art methods, NETSIM2 performs better than existing methods by incorporating gene co-functional network effectively.
- A random walk with restart-based method is developed to take both direct and indirect interactions into account.
- A standard score-based method is proposed to select the significant GO-term pairs to measure the semantic similarity.

Methods

NETSIM2 calculates the semantic similarity between two genes in three steps (see Fig. 1). First, given a gene co-functional network, it computes the relevance score between two genes based on a random walk with restart method. Second, it calculates the similarity between two GO terms by combining the information from co-functional network and GO. Finally, it selects the significant GO-term pairs to measure the similarity of two genes using a standard score-based method.

Calculating the relevance score between genes

In this step, we consider both the direct and indirect interactions in the gene co-functional network to calculate

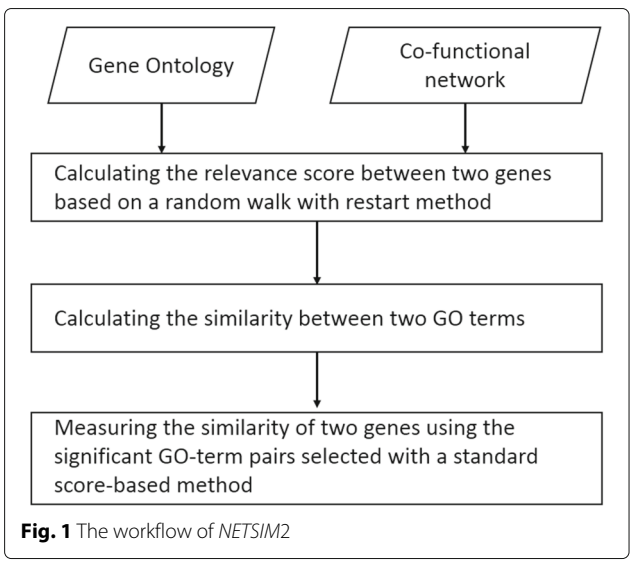

Fig. 1 The workflow of *NETSIM2*

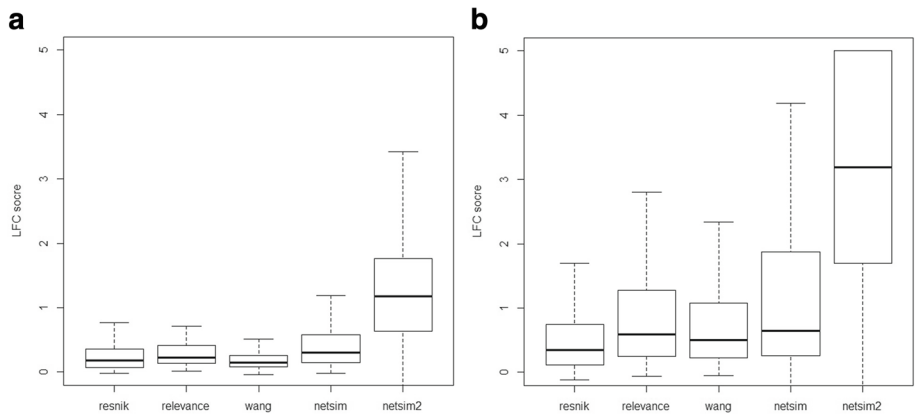

Fig. 2 Performance comparison of different measures on GO's molecular function terms in yeast (**a**) and Arabidopsis (**b**)

the relevance score between two genes. A gene network includes not only the direct interactions but also the associations between indirectly connected genes. In this step, we adopted the random walk with restart (RWR) [28] algorithm to measure the relevance score between two genes. The relevance score between genes could be represented by the stationary probability calculated by RWR. Comparing with the direct interactions, the relevance score defined by RWR can capture the global structure information of the co-functional network [29]. Furthermore, comparing with the graph distance metrics (such as shortest path), it can reveal the multi-facet relationship between two genes [30].

In RWR method, a random process begins from gene i. It iteratively transmits to neighbors of i with the probability that is based on the weights of edges. Similarly, the particle has the probability c to go back to start gene i. The association score between gene i and gene j could be defined as the stationary probability $r[i,j]$ that the iteration process will finally stop at gene j. Mathematically, given a co-functional network $N(V,E)$, the relevance scores between genes can be calculated by following steps. First, given a weight matrix M corresponding to N, a normalized weighted matrix M' was generated. Then, the RWR-based method could be described as follows.

$$\mathbf{r}_{i+1} = cM'\mathbf{r}_i + (1-c)\mathbf{e}_i \tag{1}$$

where $\mathbf{r}_i$ is a $|V| \times 1$ vector and $\mathbf{e}_i$ is a $|V| \times 1$ starting vector (the i^{th} element is 1 and others 0). $(1-c)$ is defined as the restart probability, which is between 0 and 1. Based on Equation 1, $\mathbf{r}_i$ can be defined as follows.

$$\mathbf{r}_i = (1-c)(\boldsymbol{I} - cM')^{-1}\mathbf{e}_i \tag{2}$$

After this step, we can get a matrix R, which saved the relevance scores between each pair of genes in $N(V,E)$.

Calculating the similarity between two GO terms

In this step, we calculate the similarity between two GO terms combining the information from co-function network and GO based on the method we represented in our previous work [19].

Let t_1 and t_2 be two terms. We define $D(t_1,t_2)$ as the gene set distance to compute the similarity between sets of genes annotated by t_1 and t_2. $D(t_1,t_2)$ is defined as:

$$D(t_1,t_2) = \frac{\sum_{g_i \in G_1} \prod_{g_j \in G_2} d_{ij} + \sum_{g_i \in G_2} \prod_{g_j \in G_1} d_{ij}}{2|G_1 \cup G_2| - \sum_{g_i \in G_1} \prod_{g_j \in G_2} d_{ij} - \sum_{g_i \in G_2} \prod_{g_j \in G_1} d_{ij}} \tag{3}$$

where G_1 and G_2 are the gene sets annotated by t_1 and t_2 respectively. d_{ij} is the distance score between two genes, $d_{ij} = 1 - R_{ij}$. R_{ij} is the relevance score between gene i and j calculated by RWR-based method. The gene set distances of all term pairs are normalized between 0 and 1.

Then, we calculate the similarity between two terms based on a "path-constrained annotation", labeled as U. In traditional lowest common ancestor (LCA)-based methods, all the descendants of LCA are considered. The "path-constrained annotation" method only uses the terms that are the most relevant to the compared terms. The set of relevant terms includes three parts: the gene set annotated by term t_1 and t_2, and the gene set annotated by the common parent p of t_1 and t_2 and its descendant terms that are on the paths from t_1 or t_2 to p.

Table 1 The LFC scores of five methods for the molecular function category on yeast data

Method	Resnik	Relevance	Wang	NETSIM	NETSIM2
25%	0.07	0.14	0.08	0.15	0.64
50%	0.18	0.23	0.15	0.31	1.18
75%	0.36	0.42	0.25	0.58	1.76

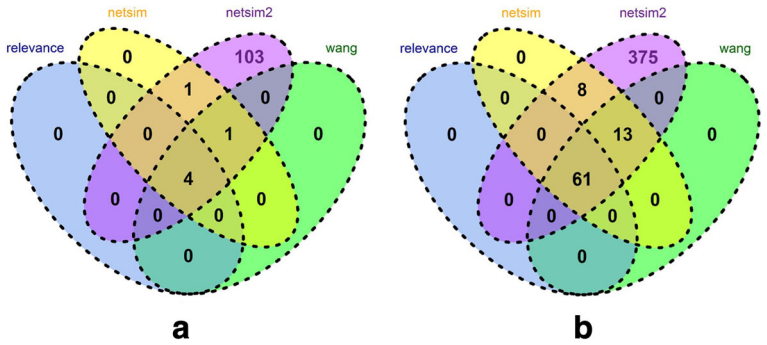

Fig. 3 Number of ECs for which NETSIM2, NETSIM, Wang and Relevance measures performed the best for yeast (**a**) and Arabidopsis (**b**) based on molecular function terms

Let t_1 and t_2 be two GO terms and p be their common ancestor. Then, the similarity between t_1 and t_2 is defined based on the equation proposed in our previous work [19].

$$S(t_1,t_2) = \frac{2log|G| - 2logf(t_1,t_2,p)}{2log|G| - (log|G_1| + log|G_2|)} \times \left(1 - \frac{h(t_1,t_2)}{|G|} \times \frac{G_p}{G}\right) \quad (4)$$

where G_p (or G) is the gene set annotated by common ancestor term p (or root term) and its descendants. In the equation, $f(t_1,t_2,p)$ calculates the similarity based on the path-constrained annotations, and is defined as follows.

$$f(t_1,t_2,p) = D(t_1,t_2)^2 \times |U(t_1,t_2,p)| + \left(1 - D(t_1,t_2)^2\right) \times \sqrt{|G_1| \times |G_2|} \quad (5)$$

$h(t_1,t_2)$ measures the specificity of the common parent, and is defined as follows.

$$h(t_1,t_2) = D(t_1,t_2)^2 \times |G| + \left(1 - D(t_1,t_2)^2\right) \times max(|G_1|,|G_2|) \quad (6)$$

In Eq. 4, the left part measures the distance from term t_1 and t_2 to p, and the right part calculates the distance from p to root. It is noted that we selected the highest score as the similarity between t_1 and t_2, if there are more than one lowest common ancestor.

Measuring the similarity of two genes

Considering both the direct and indirect interactions in the gene co-functional network may import noise information. In this step, to decrease the noise, we select the significant term pairs to calculate the gene similarities.

Let g_i and g_j be two genes. T_i and T_j are the annotation sets of g_i and g_j. Let T_G be the set of all terms contained in a GO category. Given a term t, we calculate similarities between t and each term in T_G/t, saved as S_t. Let t' be a term in T_G/t. The standard score of similarity $z_{t,t'}$ is defined as follows.

$$z_{t,t'} = \frac{S(t,t') - \mu_t}{\sigma_t} \quad (7)$$

where μ_t is the mean of the S_t and σ_t is the standard deviation of S_t. If $|z_(t,t')|$ is larger than 1.6 ($p-value$ is less than 0.05), pair (t,t') is considered as a significant term pair.

The gene similarity are calculated as follows:

$$GeneSim(g_i,g_j) = \frac{\sum_{t\in T_i} Sim\left(t,T_j'\right) + \sum_{t\in T_j} Sim\left(t,T_i'\right)}{|T_i| + |T_j|} \quad (8)$$

where T_j' (T_i') is the term set selected from T_j (T_i). To test the similarity between term $t \in T_i$ and term set T_j, we first select a term set T_j' from T_j. Based on the standard score, given term t, we can select two significant sets from T_j: $T'_{th} = \left\{t' \middle| \left(z_{t,t'} > 1.6\right)\right\}$ or $T'_{tl} = \left\{t' \middle| \left(z_{t,t'} < -1.6\right)\right\}$. If $\left|T'_{th}\right| > \left|T'_{tl}\right|$, then $T_j' = T'_{th}$, else $T_j' = T'_{tl}$. T_i' is obtained in the similar way. Choosing the significant terms to calculate the gene similarity can decrease the noise information. Each term $t \in T_i(T_j)$ can find at least a term in $T_j(T_i)$ to make a significant term pair. For each $t \in T_x$, $Sim\left(t,T_y'\right) = max_{t_y\in T_y'} S\left(t,t_y\right)$.

Results and discussion

Data preparation

We downloaded the GO structure and annotations from GO website in Dec. 2016 (www.geneontology.org). In our

Table 2 The LFC scores of five methods for the molecular function category on Arabidopsis data

Method	Resnik	Relevance	Wang	NETSIM	NETSIM2
25%	0.12	0.25	0.22	0.26	1.69
50%	0.35	0.59	0.51	0.65	3.19
75%	0.75	1.27	1.07	1.87	5

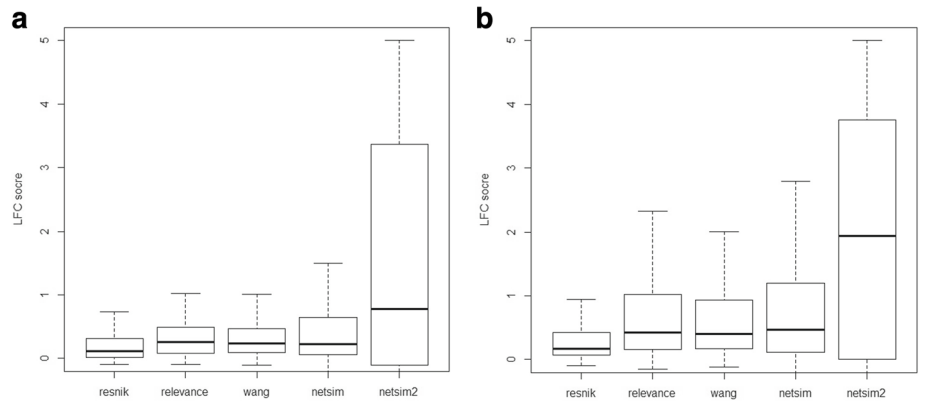

Fig. 4 Performance comparison on LFC scores of similarity measures on GO's biological process in yeast (**a**) and Arabidopsis (**b**)

work, only the is-a and part-of relationships were used. We used gene associations included in YeastNet [31]and AraNet [32] for evaluation test on yeast and arabidopsis respectively. The EC group of Yeast and Arabidopsis were downloaded from http://www.yeastgenome.org/ and http://ftp.plantcyc.org/Pathways respectively.

Performance evaluation criteria

NETSIM2 is evaluated based on the EC number (Enzyme Commission) group information, which has been used in previous research [18]. The idea is that genes that are labeled by the same EC number have the similar function. Genes are grouped to different categories based on their EC numbers (full four digits). Then, we test whether the genes in the same category have higher similarity than genes in different categories. Mathematically, we use the logged fold change (LFC) measure [18] for quantitative evaluation. The LFC score of EC number e_i is calculated as follows:

$$LFC(e_i) = \frac{1}{|EC|} \times \sum_{e_j \in EC; G(e_j) \cap G(e_i) = \emptyset} \frac{\sum_{g \in G(e_i)} diff_g(e_i, e_j)}{|G(e_i)|} \tag{9}$$

where $G(e_i)$ is gene set that includes genes labeled by e_i; EC is a set of ECs satisfying that no annotated genes is included in e_i ($G(e_j) \cap G(e_i) = \emptyset$); and $diff_g(e_i, e_j)$ is defined as:

$$diff_g(e_i, e_j) = \ln \frac{|G(e_i)| \times \sum_{g' \in G(e_j)} \left(1 - GeneSim(g, g') + c\right)}{|G(e_j)| \times \sum_{g^* \in G(e_i)} \left(1 - GeneSim(g, g^*) + c\right)} \tag{10}$$

$G(e_i)$ is the gene set of e_i without g; $G(e_j)$ is the gene set of e_j; where c is a Laplacian smoothing parameter; g is a gene assigned to e_i. $GeneSim(g, g')$ and $GeneSim(g, g^*)$ are defined in Eq. 8. Equation 10 measures the difference between the inter-EC distance and intra-EC distance.

Performance evaluation on molecular function category

The performance of *NETSIM*2 was evaluated by comparing the GO-based similarity between genes in different EC categories and same category. In this subsection, the gene similarities are calculated based on molecular function category and co-functional network. We used LFC score as a criteria to compare five measures (Resnik [33], Relevance [12], Wang [13], NETSIM [19] and *NETSIM*2) on both yeast and arabidopsis data.

NETSIM2 performed the best in all tests. In yeast, the LFC score of NETSIM2 was the highest in all tested measures (Fig. 2a, Table 1). Specifically, the median, 75th and 25th percentile value of LFC scores of NETSIM2 on yeast were 1.18, 1.76 and 0.64, significantly higher than the other measures. Interestingly, the performance of NETSIM2 was significantly higher than our previous measure NETSIM, indicating that considering the global structure of co-functional network can improve the performance. Comparing the LFC scores on each EC group using NETSIM2, NETSIM, Relevance and Wang measure (top four measures), the result shows that NETSIM2 has the highest LFC score in all 109 ECs, while NETSIM, Relevance and Wang measure has the highest LFC score in 6, 4 and 5 ECs only (Fig. 3a).

Similarly, the LFC score of NETSIM2 was the highest in all evaluated measures in arabidopsis data (Fig. 2b,

Table 3 The LFC scores of five methods for the biological process category on yeast data

Method	Resnik	Relevance	Wang	NETSIM	NETSIM2
25%	0.01	0.08	0.10	0.06	0.11
50%	0.12	0.26	0.24	0.23	0.78
75%	0.31	0.49	0.47	0.64	3.37

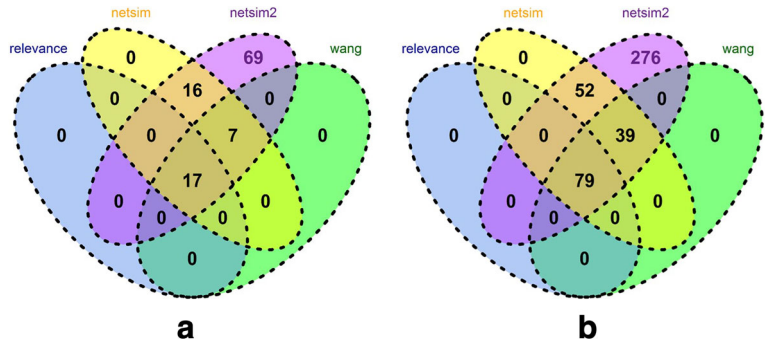

Fig. 5 Number of ECs for which NETSIM2, NETSIM, Wang and Relevance measures performed the best for yeast (**a**) and Arabidopsis (**b**) based on biological process terms

Table 2). Figure 2b shows that NETSIM2 performed significantly better than other measurements in arabidopsis data. Specifically, the 75th percentile of NETSIM2 is 5, which is the highest in all tested methods. The score of NETSIM, Relevance, Wang and Resnik measure are 1.87, 1.27, 1.07 and 0.75 respectively. The 50th percentile of NETSIM2 is 3.19, which is about 5 times of the second best measure NETSIM (0.65). Comparing the LFC scores on each EC group using NETSIM2, NETSIM, Relevance and Wang measure (top four measures), the result shows that NETSIM2 got the highest LFC score in all 457 ECs, while the number for NETSIM, Relevance and Wang measure were 82, 61 and 74 respectively (Fig. 3b). It is noted that we set the higher bound of the LFC scores as 5.

All these results indicate that NETSIM2 can improve the precision of semantic similarity measurement on molecular function category by incorporating co-function network effectively.

Performance evaluation on biological process category

In this subsection, we evaluated NETSIM2 on the biological process category. The same LFC score (Eq. 9) were used in the performance evaluation. We also evaluated NETSIM2 on both yeast and arabidopsis data.

Overall, NETSIM2 performed better than other four measures (NETSIM, Wang, Relevance and Resnik). In yeast, the 75th and median percentile of LFC scores were significant higher than other measures (Fig. 4a, Table 3), indicating that considering the global structure of co-function network and noise decrease can improve the overall performance. Specifically, the 75th percentile of LFC scores is 3.37, while the values of other measures are all less than 1 (0.64, 0.47, 0.49 and 0.31 for NETSIM, Wang, Relevance and Resnik respectively). Comparing the LFC scores on each EC group using NETSIM2, NETSIM, Relevance and Wang measure (top four measures), the result shows that NETSIM2 has the highest LFC score in all 109 ECs, while NETSIM, Relevance and Wang measure have the highest LFC score in 40, 17 and 24 ECs respectively (Fig. 5a).

Similarly, NETSIM2 performs the best in all tested measures based on biological process category in arabidopsis data (Fig. 4b, Table 4). The median and 75th percentile of LFC scores for NETSIM2 are 1.94 and 3.75, which are significant higher than the second-best measure NETSIM, which are 0.47 and 1.19 respectively (Fig. 4b and Table 4). In addition, Only NETSIM2 performs best in 276 ECs in the testing set arabidopsis ECs (Fig. 5b). For all ECs, NETSIM2 performs best, while the second best method performs best on 170 ECs.

In evaluation on both molecular function and biological process category, NETSIM2 improves more on arabidopsis data than yeast data. The reason may be that yeast data in GO is more complete than arabidopsis data. Therefore, incorporating co-functional network can improve the performance significantly on the arabidopsis data.

Conclusions

Gene Ontology (GO) is one of the most popular bioinformatics resources used to describe the properties of genes and gene products. Calculating GO-based gene functional similarity has been widely used in multiple research areas. However, the low-quality similarity may result from the incomplete information of GO and the limited amount of

Table 4 The LFC scores of five methods for the biological process category on Arabidopsis data

Method	Resnik	Relevance	Wang	NETSIM	NETSIM2
25%	0.07	0.15	0.17	0.12	0.002
50%	0.17	0.43	0.40	0.47	1.94
75%	0.42	1.03	0.91	1.19	3.75

annotations in GO. A recent measure, named NETSIM, addresses these problems by considering both gene-gene associations, GO DAG and annotations. Unfortunately, only the local association information in gene co-function network was used, since NETSIM only considers the direct link in the network.

In this paper, we proposed a novel network-based method, named NETSIM2, by considering the global structure of the co-functional network with a RWR-based method, and by selecting the significant term pairs to decrease the noise information. NETSIM2 includes three steps: firstly, given a gene co-functional network, the relevance scores between two genes are calculated based on a random walk with restart method; secondly, the similarity between two GO terms is calculated by combining the information from co-functional network and GO; finally, the significant GO-term pairs are selected to measure the similarity of two genes using a standard score-based method. Experimental results using ECs on both molecular function and biological process category show that NETSIM2 performs the best among all the measures on both yeast and Arabidopsis data set. It also shows that NETSIM2 can significantly improve the performance of semantic similarity measurement especially on the incomplete species. It is note that we have proposed NETSIM in our previous work to incorporate co-function network to GO-based semantic similarities, which can be considered as a simplified case of NETSIM2.

Acknowledgements
This work was supported by National Natural Science Foundation of China (Grant No. 61702421), Natural Science Basic Research Plan in Shaanxi Province of China (No. 2017JQ6047), China Postdoctoral Science Foundation (No. 2017M610651), the Fundamental Research Funds for the Central Universities (Grant No. 3102016QD003), National Natural Science Foundation of China (Grant No. 61702421 and 61332014).

Funding
Publication of this article was funded by Northwestern Polytechnical University.

Availability of data and materials
The data can be downloaded following the instruction in the "Data preparation" subsection.

About this supplement
This article has been published as part of *BMC Systems Biology* Volume 12 Supplement 2, 2018: Proceedings of the 28th International Conference on Genome Informatics: systems biology. The full contents of the supplement are available online at https://bmcsystbiol.biomedcentral.com/articles/supplements/volume-12-supplement-2.

Authors' contributions
JP and XS designed the web tool framework; JP and XZ implemented the method; JP wrote this manuscript; SL, WH, JL and QL helped design the input interface. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]School of Computer Science, Northwestern Polytechnical University, Xi'an, China. [2]Key Laboratory of Big Data Storage and Management, Northwestern Polytechnical University, Ministry of Industry and Information Technology, Xi'an, China. [3]Centre for Multidisciplinary Convergence Computing (CMCC), School of Computer Science, Northwestern Polytechnical University, Xi'an, China.

Published: 19 March 2018

References

1. Consortium GO, et al. Expansion of the gene ontology knowledgebase and resources. Nucleic Acids Res. 2017;45(D1):331–38.
2. Vafaee F, Rosu D, Broackes-Carter F, Jurisica I. Novel semantic similarity measure improves an integrative approach to predicting gene functional associations. BMC Syst Biol. 2013;7(1):22.
3. Cheng L, Sun J, Xu W, Dong L, Hu Y, Zhou M. Oahg: an integrated resource for annotating human genes with multi-level ontologies. Sci Rep. 2016;6:34820.
4. Peng J, Bai K, Shang X, Wang G, Xue H, Jin S, Cheng L, Wang Y, Chen J. Predicting disease-related genes using integrated biomedical networks. BMC Genomics. 2017;18(1):1043.
5. Peng J, Wang T, Wang J, Wang Y, Chen J. Extending gene ontology with gene association networks. Bioinformatics. 2015;32(8):1185–94.
6. Díaz-Montaña JJ, Díaz-Díaz N, Gómez-Vela F. Gfd-net: A novel semantic similarity methodology for the analysis of gene networks. J Biomed Inform. 2017;68:71–82.
7. Yu G, Fu G, Wang J, Zhu H. Predicting protein function via semantic integration of multiple networks. IEEE/ACM Trans Comput Biol Bioinforma. 2016;13(2):220–32.
8. Nehrt NL, Clark WT, Radivojac P, Hahn MW. Testing the ortholog conjecture with comparative functional genomic data from mammals. PLoS Comput Biol. 2011;7(6):1002073.
9. Peng J, Li H, Liu Y, Juan L, Jiang Q, Wang Y, Chen J. Intego2: a web tool for measuring and visualizing gene semantic similarities using gene ontology. BMC Genomics. 2016;17(5):530.
10. Yang Y, Xu Z, Song D. Missing value imputation for microRNA expression data by using a GO-based similarity measure. BMC bioinformatics. 2016;17(1):S10. BioMed Central.
11. Peng J, Lu J, Shang X, Chen J. Identifying consistent disease subnetworks using dnet. Methods. 2017;131:104–10.
12. Schlicker A, Domingues FS, Rahnenführer J, Lengauer T. A new measure for functional similarity of gene products based on gene ontology. BMC Bioinformatics. 2006;7(1):302.
13. Wang JZ, Du Z, Payattakool R, Yu PS, Chen C-F. A new method to measure the semantic similarity of go terms. Bioinformatics. 2007;23(10):1274–81.
14. Pesquita C, Faria D, Falcao AO, Lord P, Couto FM. Semantic similarity in biomedical ontologies. PLoS Comput Biol. 2009;5(7):1000443.
15. Yang H, Nepusz T, Paccanaro A. Improving go semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. Bioinformatics. 2012;28(10):1383–9.
16. Teng Z, Guo M, Liu X, Dai Q, Wang C, Xuan P. Measuring gene functional similarity based on group-wise comparison of go terms. Bioinformatics. 2013;29(11):1424–32.
17. Peng J, Wang Y, Chen J. Towards integrative gene functional similarity measurement. BMC Bioinformatics. 2014;15(2):5.
18. Peng J, Li H, Jiang Q, Wang Y, Chen J. An integrative approach for measuring semantic similarities using gene ontology. BMC Syst Biol. 2014;8(5):8.
19. Peng J, Uygun S, Kim T, Wang Y, Rhee SY, Chen J. Measuring semantic similarities by combining gene ontology annotations and gene co-function networks. BMC Bioinformatics. 2015;16(1):44.

20. Mazandu GK, Chimusa ER, Mulder NJ. Gene ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery. Brief Bioinforma. 2016;18(5):886–901.
21. Zhang S-B, Lai J-H. An integrated information-based similarity measurement of gene ontology terms. Comput Sci Inf Syst. 2015;12(4): 1235–53.
22. Peng J, Xue H, Shao Y, Shang X, Wang Y, Chen J. A novel method to measure the semantic similarity of hpo terms. Int J Data Min Bioinforma. 2017;17(2):173–88.
23. Peng J, Wang H, Lu J, Hui W, Wang Y, Shang X. Identifying term relations cross different gene ontology categories. BMC Bioinformatics. 2017;18(16):573.
24. Wu H, Su Z, Mao F, Olman V, Xu Y. Prediction of functional modules based on comparative genome analysis and gene ontology application. Nucleic Acids Res. 2005;33(9):2822–37.
25. Wu X, Pang E, Lin K, Pei Z-M. Improving the measurement of semantic similarity between gene ontology terms and gene products: insights from an edge-and ic-based hybrid method. PloS ONE. 2013;8(5):66745.
26. Sevilla JL, Segura V, Podhorski A, Guruceaga E, Mato JM, Martinez-Cruz LA, Corrales FJ, Rubio A. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB). 2005;2(4):330–8.
27. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, et al. The arabidopsis information resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2011;40(D1):D1202–10.
28. Tong H, Faloutsos C, Pan J-Y. Fast Random Walk with Restart and Its Applications. In: Proceedings of the Sixth International Conference on Data Mining, ICDM '06. Washington: IEEE Computer Society; 2006. p. 613–22. https://doi.org/10.1109/ICDM.2006.70.
29. He J, Li M, Zhang H-J, Tong H, Zhang C. Manifold-ranking Based Image Retrieval. In: Proceedings of the 12th Annual ACM International Conference on Multimedia, MULTIMEDIA '04. New York: ACM; 2004. p. 9–16. http://doi.acm.org/10.1145/1027527.1027531.
30. Tong H, Faloutsos C. Center-piece Subgraphs: Problem Definition and Fast Solutions. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06. New York: ACM; 2006. p. 404–13. http://doi.acm.org/10.1145/1150402.1150448.
31. Lee I, Li Z, Marcotte EM. An improved, bias-reduced probabilistic functional gene network of baker's yeast, saccharomyces cerevisiae. PloS ONE. 2007;2(10):988.
32. Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY. Rational association of genes with traits using a genome-scale gene network for arabidopsis thaliana. Nat Biotechnol. 2010;28(2):149–56.
33. Resnik P, et al. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J Artif Intell Res (JAIR). 1999;11:95–130.