

RESEARCH

Open Access



# Measuring phenotype-phenotype similarity through the interactome

Jiajie Peng, Weiwei Hui and Xuequn Shang\*

From Biological Ontologies and Knowledge bases workshop at IEEE BIBM 2017  
Kansas City, MO, USA. 14 November 2017

## Abstract

**Background:** Recently, measuring phenotype similarity began to play an important role in disease diagnosis. Researchers have begun to pay attention to develop phenotype similarity measurement. However, existing methods ignore the interactions between phenotype-associated proteins, which may lead to inaccurate phenotype similarity.

**Results:** We proposed a network-based method *PhenoNet* to calculate the similarity between phenotypes. We localized phenotypes in the network and calculated the similarity between phenotype-associated modules by modeling both the inter- and intra-similarity.

**Conclusions:** *PhenoNet* was evaluated on two independent evaluation datasets: gene ontology and gene expression data. The result shows that *PhenoNet* performs better than the state-of-art methods on all evaluation tests.

**Keywords:** Phenotype relationships, Interactome, Human phenotype ontology

## Background

Recently, advances in next generation sequencing (NGS) have significantly improved the Mendelian disease diagnosis [1–4]. However, disease diagnosis only using sequence-based approach is still challenging, since lots of diseases have complex phenotypes and high genetic heterogeneity. It is difficult to reveal the relationship between genetic features and complex patient phenotypic features [5].

In medical contexts, “phenotype” often refers to the deviation from normal morphology, physiology, or behavior [6]. Phenotype carries the biologically meaningful information [7, 8]. Disease is usually the result of congenital or acquired mutations [9, 10], which causes the phenotypes of diseases [11]. Therefore, phenotype plays an important role in disease diagnosis process. Clinical practice and medical research based on phenotype analysis have drawn great attention in recent years [12–14]. Measuring phenotype similarity became one of

the key components in disease diagnosis and understanding the disease mechanism [15].

In the past few years, a lot of computational methods have been proposed to calculate phenotype similarity [7, 11, 16–24]. These approaches can be loosely grouped into two categories: text mining-based method and ontology-based method. Text mining-based methods measure the phenotype similarity by comparing texts describing phenotypes. These methods extract features from the descriptions and construct machine learning models [25, 26] based on these features [7, 8]. Driel et al. classified more than 5000 phenotypes based on texts from the OMIM database [7]. They found the similarity between phenotypes are positively correlated with protein sequence similarity and protein-protein interaction. However, text data always results to ambiguous meanings. For example, different words may represent the same meaning.

To avoid ambiguous representation, a unified vocabulary system was developed to describe phenotype, named Human Phenotype Ontology (HPO). HPO is widely used to describe human phenotypic abnormalities using a structured, controlled and unified vocabularies, which are constructed as a directed acyclic graph (DAG). The

\*Correspondence: [shang@nwpu.edu.cn](mailto:shang@nwpu.edu.cn)

<sup>1</sup>School of Computer Science, Northwestern Polytechnical University, Xi’an, China

ontology-based method measures the phenotype similarity based on the HPO. HPO-based semantic similarity has been used to quantify the phenotypic similarity between patient symptoms and known phenotypes related to a gene [27, 28]. This type of method is based on hierarchical structure of HPO and the annotations of phenotypes [11, 29–33]. Phomizer and Masino et al. calculated the similarity between phenotypes based on the information content (IC) of their lowest common-ancestor in HPO [27, 28]. Given a term  $t$ , its IC is calculated as:  $IC(t) = -\log \frac{|D_t|}{D}$ , where  $D_t$  and  $D$  represent annotation set of  $t$  and all annotations involved in HPO respectively. Performance evaluation shows that this method performs better than the term matching-based approach that ignores the semantic relation between terms. However, this method may be hindered by the noises in the patient phenotype data [28]. A PageRank-based method, named *PhenoSim*, was proposed to model the noises in the patient phenotype data set before applying the ontology-based method [23]. The evaluation test shows that *PhenoSim* performs better than the IC-based method. However, the ontology-based method relies on the hierarchical structure of HPO, which makes it hard to distinguish the similarities between terms that have the same lowest common ancestor. For example, let  $A$ ,  $B$  and  $C$  be terms with the same common ancestor. The ontology-based method can not measure whether similarity of  $A$  and  $B$  are higher than  $B$  and  $C$ . Furthermore, aforementioned existing methods ignores the interactions between the proteins annotated by the phenotypes, which is critical to understand the molecular basis of phenotypes.

Recently, network-based method has been proposed to measure the similarity between diseases [20, 34–36]. The similarity between two diseases is measured by exploring the relationships between disease-associated proteins within the interactome. Specifically, given disease  $d_a$ ,  $d_b$  and their associated gene set  $G_a$  and  $G_b$ , the similarity between  $d_a$  and  $d_b$  is calculated as follows.

$$s_{ab} = h_{ab} - \frac{h_{aa} + h_{bb}}{2} \quad (1)$$

where  $h_{ab}$  is the average of shortest distances of protein pairs across  $G_a$  and  $G_b$ .  $h_{aa}$  ( $h_{bb}$ ) is the average of shortest distances in  $d_a$  ( $d_b$ ). The evaluation test shows that the network-based similarities have high correlation with gene co-expression patterns, symptom similarities and gene ontology-based similarities [20].

Unfortunately, to our knowledge, the network-based model has not been used to measure the similarity between phenotypes. To overcome the disadvantages in aforementioned phenotype similarity measurements, we present a network-based method, called *PhenoNet*, to calculate the phenotype similarity by comparing the phenotype-associated modules in the protein-protein

interaction network. We localize the phenotypes in the network and propose a network-based method considering both intra- and inter-similarity of the phenotype-associated modules, which can effectively use the rich information in the network to measure the phenotype similarity. Comparing with the existing approaches, the contributions of our work are listed as follows:

- Our work indicates that phenotypes can be represented by the network modules.
- To the best of our knowledge, *PhenoNet* is the first phenotype similarity measurement based on the interactome.
- We proposed a novel method to calculate similarity between phenotypes considering both the inter- and intra-similarity of the phenotype-associated module in the network.

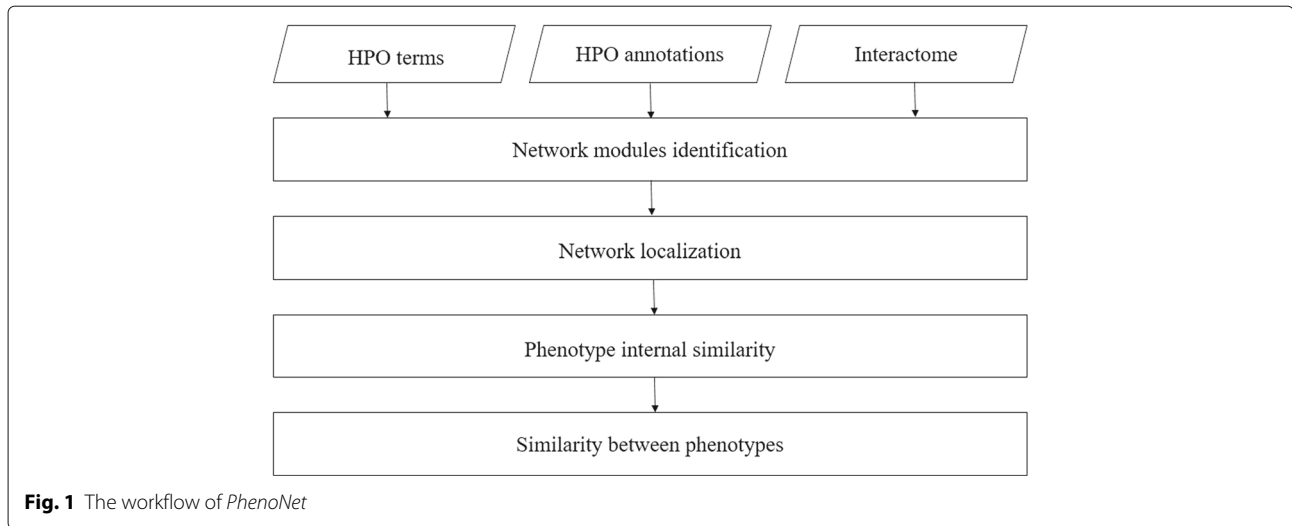
## Methods

We propose a new network-based method called *PhenoNet* to measure the relationships between phenotypes based on the biological network. Our model includes four parts. Firstly, for each given phenotype  $p$ , we identify the corresponding module in the given biological network, labeled as  $n_p$ . Secondly, statistical method is used to test whether phenotypes could be represented by network modules. Thirdly, for each network module  $n_p$ , we first calculate the internal similarity by modeling the relationships among nodes in  $n_p$ . Finally, given two phenotypes  $p_1$  and  $p_2$ , their relationship is measured based on corresponding network module  $n_{p_1}$  and  $n_{p_2}$ . The diagram of the whole process is shown in Fig. 1.

### Network modules identification

In order to measure the relationship between two phenotypes based on a biological network  $N$ , the first step is to identify network modules in  $N$  to represent phenotypes. In this step, we will generate a set of network modules corresponding to a set of phenotypes. In general, network module means a group of nodes that associate with a specific function, disease or phenotype *etc*. Given a biological network  $N$ , our goal is to identify a network module corresponding to a phenotype  $p$ .

HPO is constructed as a directed acyclic graph (DAG), in which each term represents a phenotype. Each phenotype term  $p$  is related to a set of proteins  $G_p$ , which can be used to identify the phenotype module in a given biological network. Specifically, given a biological network  $N(V, E)$ , a phenotype  $p$  and a set of proteins  $G_p$  related to  $p$ , the network module of  $p$  in  $N$  is a subnetwork of  $N$ , labeled as  $N_p(V', E')$ .  $V'$  is the intersection set of  $V$  and  $G_p$ .  $E'$  is a subset of  $E$ , which connects the nodes in  $E'$ .



**Network localization**

Given a network module, we test whether proteins in a network module can agglomerate in specific interactome neighborhoods. If a network module of a phenotype is a highly interconnected group of proteins, the phenotype could be represented by a network module. Then, we could measure the relations between phenotypes by comparing network modules that represent phenotypes.

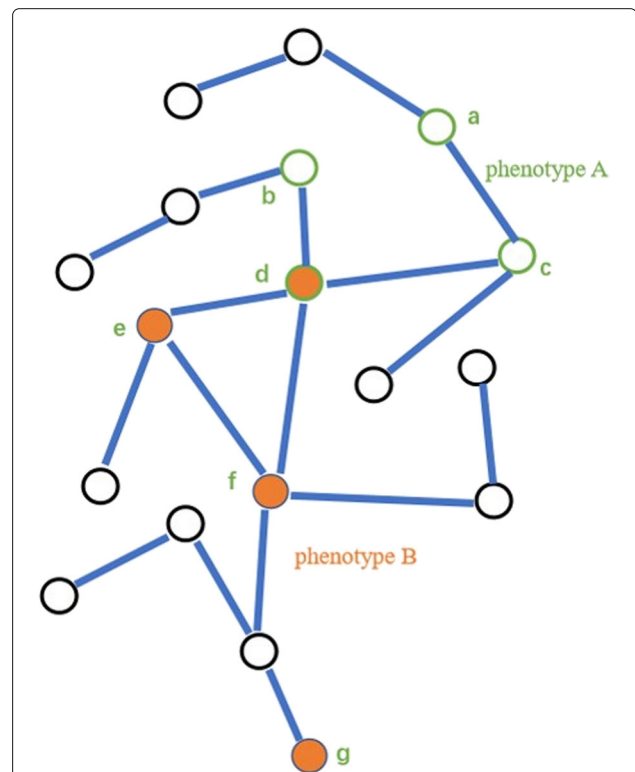
We used two metrics to quantify the localization of the identified network module. One is the module size. Module size is defined as the number of proteins contained in the largest connected subnetwork of a network module. The other is the shortest distance. For each protein, the shortest distance is defined as the distance to the closest protein in the same network module. Given a network module  $N_A$  of phenotype  $A$ , the shortest distance is defined as follows.

$$d_A = \frac{\sum_{v \in N_A} d_v}{|N_A|} \tag{2}$$

where  $d_v$  is the shortest distance of a protein  $v$  in the network module  $N_A$ . In the illustrative example in Fig. 2, the module sizes of network modules corresponding to phenotype  $A$  and  $B$  are 4 and 3 respectively. The shortest distances of network modules corresponding to phenotype  $A$  and  $B$  are 4 and 5 respectively.

To test whether the proteins in a network module agglomerate together, we compare the module size and shortest distance of a network module with a random control. Given a phenotype  $A$  and its corresponding network module  $N_A$ , we randomize the phenotype associations of proteins and generate a random network module  $N'_A$  on the given PPI network.  $N'_A$  includes the same number of proteins as  $N_A$ . Given a phenotype, we repeat this procedure 100,000 times to obtain 100,000 network modules corresponding to  $A$ . Based on the randomly generated

network modules, we can calculate the statistical significance of the real data. For both module size and shortest distance, we estimate the mean values and standard deviations. Given a network module  $N_A$ , the  $z$ -score of module size is calculated as follows.



**Fig. 2** An illustration example of protein-protein interaction network. In this network, network module of phenotype  $A$  consists of the nodes with green circle. Network module of phenotype  $B$  consists of the nodes filled with orange. Specifically, phenotype  $B$  is split into two components

$$z_A^{size} = \frac{s_A - \mu_1}{\sigma_1} \tag{3}$$

where  $s_A$  is the module size of  $N_A$ .  $\mu_1$  and  $\sigma_1$  represent the mean value and standard deviation of module size based on the randomly generated network modules.

Similarly, the  $z$  - score of shortest distance is calculated as follows.

$$z_A^{dis} = \frac{d_{N_A} - \mu_2}{\sigma_2} \tag{4}$$

where  $d_{N_A}$  is the shortest path of  $N_A$ .  $\mu_2$  and  $\sigma_2$  represent the mean value and standard deviation of shortest path based on the randomly generated network modules.

We could also obtain the corresponding  $p$ -value for each  $z$  - score after calculating the  $z$  - score for every network module identified in the network. Therefore, we can test whether proteins in a identified network module are highly interconnected via the statistical significance of module size and shortest distance.

**Phenotype internal similarity calculation**

The similarity between two phenotypes could be measured by measuring the distance between their corresponding network modules in a network. To calculate the similarity between two network modules, we consider both the internal similarity in each network module and similarity between network modules. The idea is inspired by the traditional clustering evaluation method that both inter- and intra- similarity should be considered. Given two network modules  $N_A$  and  $N_B$ , their similarity in the network is proportional to the intra-similarity of  $N_A$  ( $N_B$ ) and the inter-similarity between  $N_A$  and  $N_B$ .

The intra-similarity of phenotype  $N_A$  is defined as the average of pairwise similarities between proteins contained in  $N_A$ . The similarity between two proteins  $i$  and  $j$  is defined as follows.

$$Sim(i, j) = \frac{1}{d_s(i, j)} \tag{5}$$

where  $d_s(i, j)$  is the length of the shortest path between  $i$  and  $j$ . Mathematically, the intra-similarity of network module  $N_A$  is defined as follows.

$$Sim_{intra}(N_A) = \frac{\sum_{i \in N_A} \sum_{j \in N_A, j \neq i} Sim(i, j)}{|N_A| \times (|N_A| - 1)} \tag{6}$$

The intra-similarity of a network module can be calculated based on Eq. 6. Intra-similarity can reflect whether a network module agglomerate together in the network.

**Similarity between phenotypes calculation**

The network-based similarity of phenotype  $A$  and  $B$  can be quantified by comparing the similarity between network module  $N_A$  and  $N_B$  corresponding to  $A$  and  $B$  respectively.

As we described previously, the similarity between two network modules is determined by both intra- and inter-similarity. The inter-similarity between phenotype  $N_A$  and  $N_B$  is calculated based on pairwise similarities between proteins in different network modules. Let  $i$  be a protein involved in  $N_A$ , the similarity between  $i$  and  $N_B$  is computed with the following equation:

$$S_{i \rightarrow N_B} = \begin{cases} \frac{\sum_{j \in N_B} Sim(i, j)}{|N_B|}, & i \neq j \\ 1, & i = j \end{cases} \tag{7}$$

where  $Sim(i, j)$  is the similarity between protein  $i$  and  $j$  (see Eq. 5),  $|N_B|$  represents the number of proteins involved in the network module  $N_B$ . Particularly, if  $i$  and  $j$  are identical, the similarity value is set as 1. Then, we can calculate the inter-similarity between two network modules  $N_A$  and  $N_B$  with following equation:

$$Sim_{inter}(N_A, N_B) = \frac{1}{2} \left( \frac{\sum_{i \in N_A} S_{i \rightarrow N_B}}{|N_A|} + \frac{\sum_{i \in N_B} S_{i \rightarrow N_A}}{|N_B|} \right) \tag{8}$$

where  $|N_A|$  and  $|N_B|$  represent the number of proteins involved in the network module  $N_A$  and  $N_B$  respectively. Equation 8 includes two parts: the average of similarities between each protein  $i \in N_A$  and  $N_B$ ; the average of similarities between each protein  $i \in N_B$  and  $N_A$ . Note that the aforementioned two parts are asymmetric. To avoid the asymmetry result, the inter-similarity of two network modules are calculated as Eq. 8.

By considering both inter- and intra- similarity, the similarity between two network module  $N_A$  and  $N_B$  corresponding to phenotype  $A$  and  $B$  is calculated as follows.

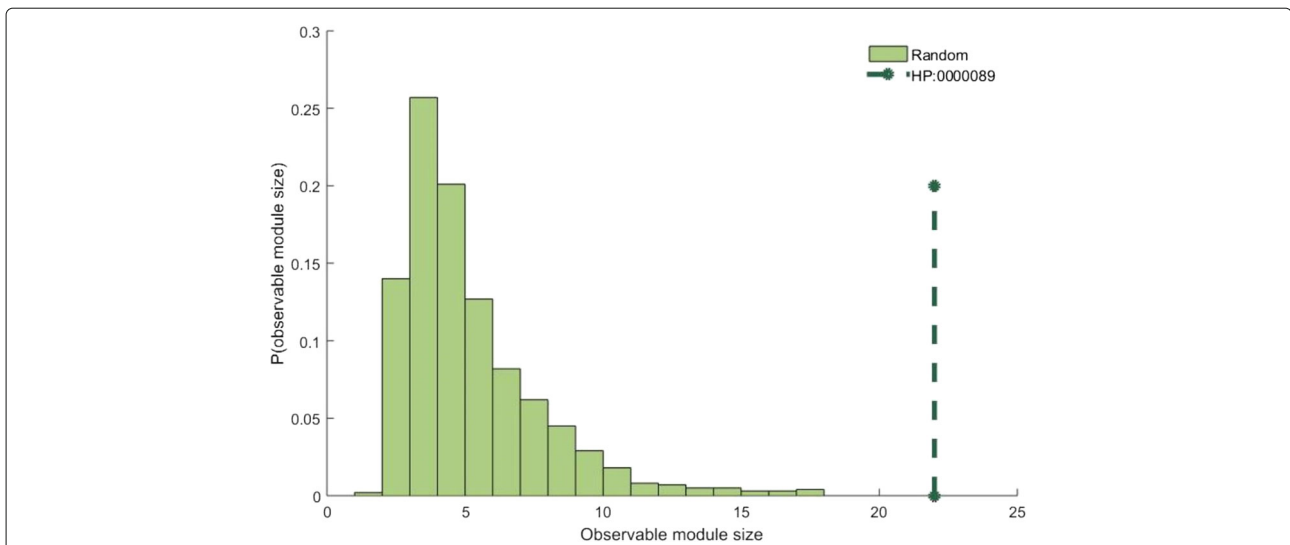
$$Sim(A, B) = Sim_{inter}(N_A, N_B) - \frac{Sim_{intra}(N_A) + Sim_{intra}(N_B)}{2} \tag{9}$$

where  $A$  and  $B$  are two phenotypes,  $N_A$  and  $N_B$  are two network modules in the PPI network corresponding to  $A$  and  $B$  respectively.

**Results and discussion**

**PhenoNet implementation and data preparation**

PhenoNet was implemented with Python 2.7 and *networkx* library. HPO data was downloaded from the HPO website in Apr. 2016 (<http://human-phenotype-ontology.github.io/downloads.html>). HPO data contains 11786 human phenotype terms. There exist relationships (is\_a) between HPO terms. We used this relationship to generate HPO term tree firstly. Then we up-propagated the HPO terms with the hierarchy of the full HPO tree. From the phenotype terms we selected the HPO terms with at least 25 genes based on the percolation theory, for the reason



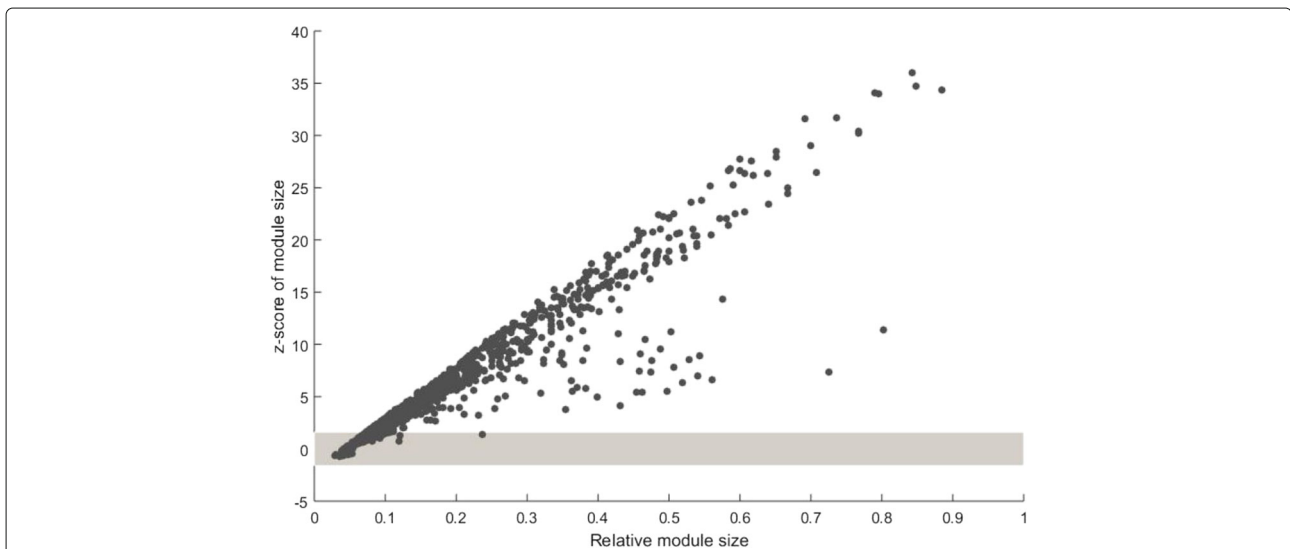
**Fig. 3** Module size distribution of the random models based on phenotype “HP:0000089”. The module size of random models is less than the phenotype

that our protein-protein interaction network is incomplete. Since our goal is to compute the similarity between different phenotypes, we deleted all the parent terms from the term sets obtained in the former step in order to eliminate the influence from parent-son relationships. Finally we used 1061 HPO terms in our experiment.

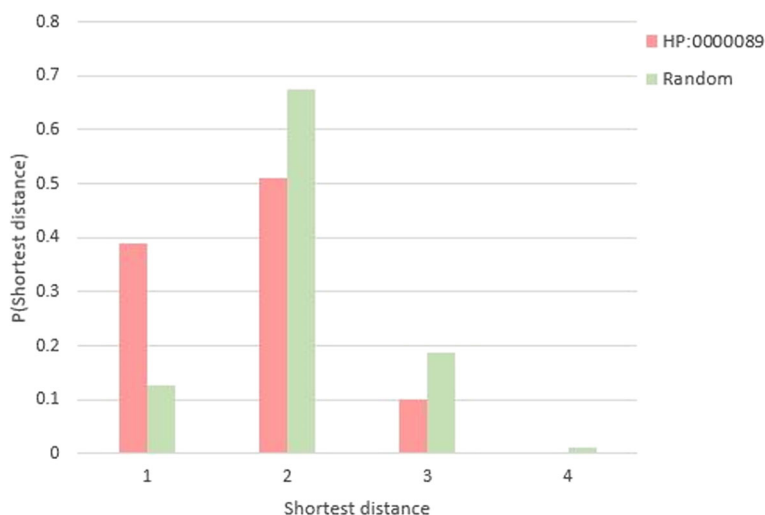
GO data was downloaded from GO website in Nov. 2016 ([www.geneontology.org/](http://www.geneontology.org/)). We only used the annotations with high confidence. Specifically, the selected annotations are associated with the following evidence codes: EXP, IDA, IMP, IGI, IEP, ISS, ISA, ISM and ISO. And we ignored the annotations with evidence code IPI to avoid

the logical cycle of the performance evaluation, since the protein-protein interaction network that used to calculate the phenotype similarity is constructed by the physical protein-protein interactions. We also removed the annotations with a non-empty “qualifier” column [37]. The annotations of a GO term includes its direct annotations and annotations of its descendants.

The gene expression data was from (<http://www.ncbi.nlm.nih.gov/geo>, downloaded Jan. 2017) [38]. We excluded nine unhealthy tissue data of the 79 tissues in the dataset [20]. The highest expression value is selected when an mRNA has multiple transcripts. After mapping mRNA



**Fig. 4** z – score of 1061 HPO terms based on their module size. Eight hundred seventy two phenotypes are significant. Points in the gray area represent the phenotypes which are not significant



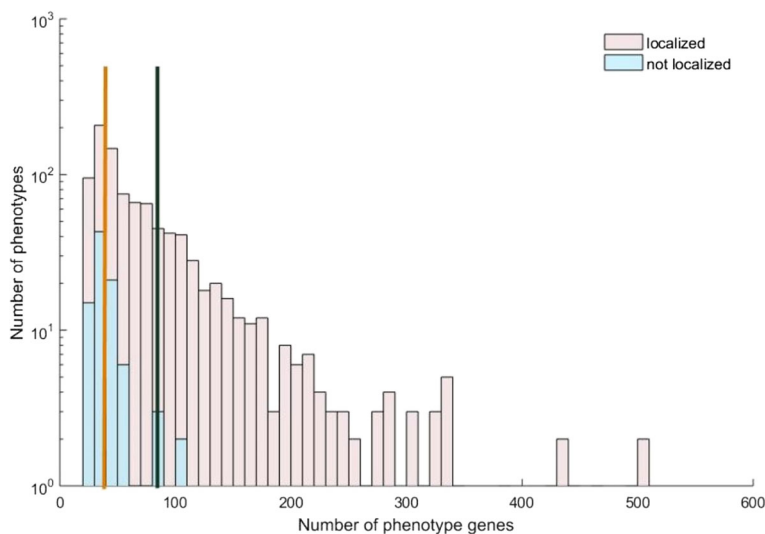
**Fig. 5** Shortest distance distribution of the random set based on HP:0000089. The shortest distance of phenotype “HP:0000089” is smaller than the randomly set. The x-axis is the shortest distance. The y-axis is the probability of each shortest distance. Let  $s$  be the size of HPO,  $n_i$  be the number of shortest distance  $i$  ( $i$  can be 1,2,3,4), then the probability is calculated by  $\frac{n_i}{1000000 * s}$

to gene, we totally obtained 2849 genes with expression value.

The interactome generated by Menche et al. was used in the following tests [20], which integrated several databases with seven types of physical interactions: regulatory interactions [39], binary interactions [40–44], literature curated interactions [45–47], metabolic enzyme-coupled interactions [48], protein complexes [49], kinase network (kinase-substrate pairs) [50] and signaling interactions [51]. The final PPI interaction network includes 141,296 interactions between 13,460 proteins. Note that the interactions extracted from gene expression data or evolutionary considerations are not included.

**Network localization**

We used phenotype module size and the shortest distance to test whether these phenotype proteins tend to agglomerate in the protein-protein interaction network. Taking phenotype “Renal hypoplasia (HP: 0000089)” as an example, the average module size of the randomly generated models is significantly less than the real module size (the number is 22) of term HP: 0000089 (Fig. 3,  $p$ -value  $\leq 10^{-5}$ , the threshold of  $z$  - score is  $z$  - score  $\geq 1.6$ ,  $p$ -value  $\leq 0.05$ ). Overall, 872 of 1061 phenotypes are statistically significant based on module size. The  $z$  - score based on the module sizes of these phenotypes are shown in Fig. 4. It shows that the module sizes of



**Fig. 6** Distribution for  $z$  - score of 1061 HPO terms on shortest distance. The pink and blue bars represent the distribution of localized and not-localized respectively. The average numbers of proteins associated with localized and not-localized phenotypes are 82 and 40 respectively



**Table 1** The Pearson correlation coefficient and  $R^2$  score between three methods and GO-based similarity

GO term	Measure		Ontology-based	Menche	PhenoNet
BP	PCC	Median	0.58	0.82	<b>0.96</b>
		Mean	0.66	0.89	<b>0.96</b>
	$R^2$	Median	0.35	0.69	<b>0.93</b>
		Mean	0.43	0.79	<b>0.92</b>
MF	PCC	Median	0.43	0.61	<b>0.94</b>
		Mean	0.48	0.66	<b>0.95</b>
	$R^2$	Median	0.17	0.37	<b>0.87</b>
		Mean	0.23	0.44	<b>0.91</b>

The highest values were highlighted in boldface

most phenotypes are significant, indicating that proteins involved in the same phenotype tend to be connected.

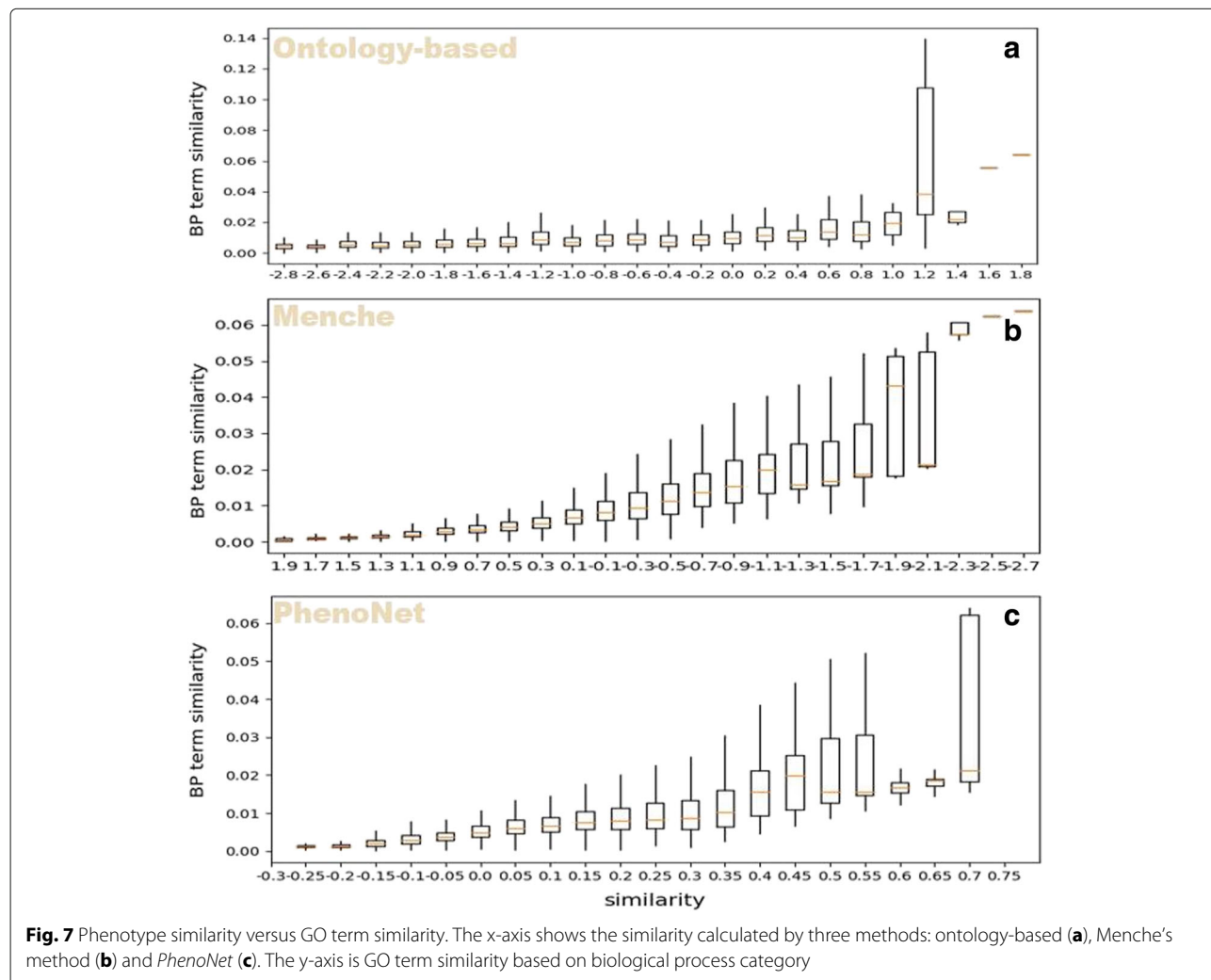
For the shortest distance, we also take phenotype “HP: 0000089” as an example. Figure 5 shows that the shortest distance of the randomly generated modules is

significantly larger than the real phenotype “HP: 0000089” ( $z$  - score = -5.51,  $p$ -value  $\leq 10^{-5}$ ). The results indicate that proteins associated with one phenotype are usually a group of interconnected nodes in the network instead of dispersing across the whole network. Overall, 969 of 1061 phenotypes have significantly shorter distances compared to the random distribution (Fig. 6). The black and yellow line in the figure show the average number of phenotype proteins of the well-localized phenotypes and not-localized phenotypes (82 and 40 respectively).

In summary, the proteins involved in a phenotype have shown a statistically significant tendency to agglomerate together in the network based on both module size and shortest distance.

**Performance evaluation on gene ontology dataset**

To evaluate the performance of *PhenoNet*, we employed GO-based similarity as the independent evidence to compare the performance of *PhenoNet* with two existing



**Fig. 7** Phenotype similarity versus GO term similarity. The x-axis shows the similarity calculated by three methods: ontology-based (a), Menche’s method (b) and *PhenoNet* (c). The y-axis is GO term similarity based on biological process category

algorithms, i.e. ontology-based method [29] and Menche’s method [20]. Ontology-based method calculated the phenotype similarity based on the IC of the lowest common ancestor of the phenotypes in HPO. Menche’s method [20] is a network-based method which was used to measure the disease similarity.

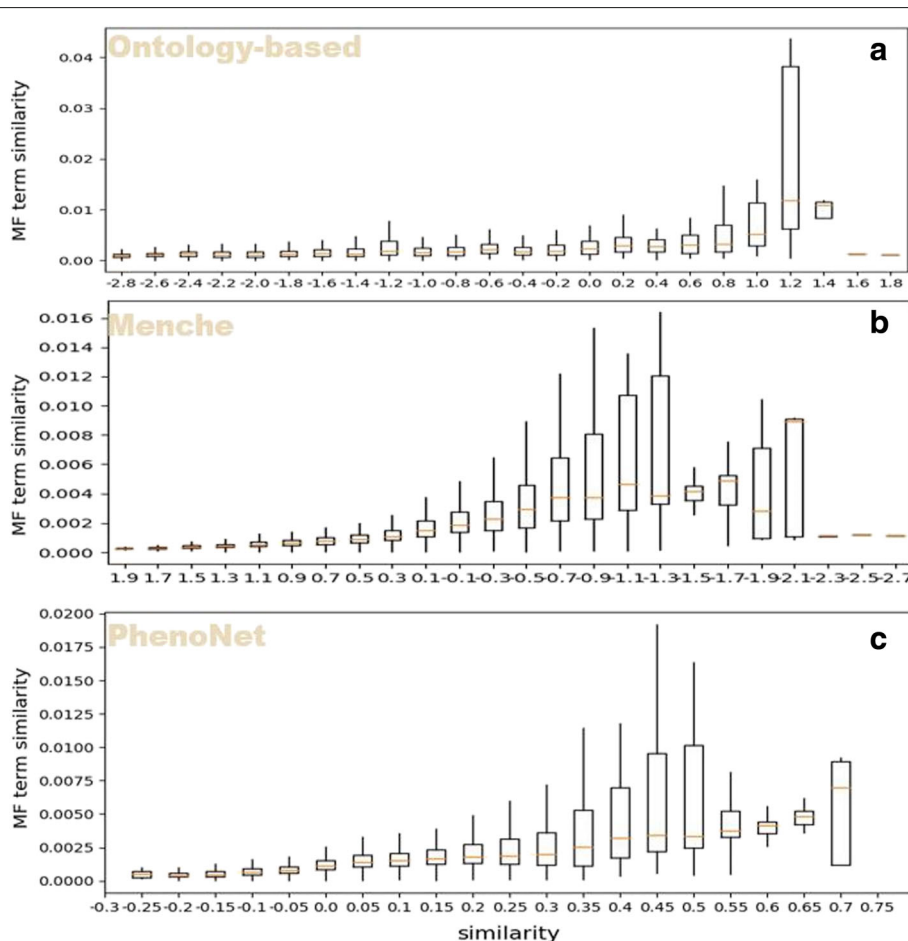
The idea is that the network-based similarity between phenotypes should be correlated with the GO-based similarity. Given two phenotypes  $p_1$  and  $p_2$ , let  $G_1$  and  $G_2$  be the protein sets associated with  $p_1$  and  $p_2$  respectively. By adopting the method used in [20], the similarity between  $p_1$  and  $p_2$  is measured by the average of all protein pairs between  $G_1$  and  $G_2$  (Equation 10).

$$Sim_{phe}(p_1, p_2) = \frac{1}{|G_1| \times |G_2|} \sum_{g_i \in G_1, g_j \in G_2} Sim_{pro}(g_i, g_j) \tag{10}$$

$Sim_{pro}(g_i, g_j)$  is the functional similarity between  $g_i$  and  $g_j$  based on GO. It is defined based on the specificity of

shared GO annotations as:  $Sim_{pro}(g_i, g_j) = \frac{2}{n_t}$ .  $n_t$  is the total number of proteins annotated to the most specific GO term  $t$  shared by  $g_i$  and  $g_j$ . Particularly,  $Sim_{pro}(g_i, g_j) = 1$  if  $g_i$  and  $g_j$  are the only two proteins annotated by a specific GO term.

With the similarity between phenotypes based on GO functional similarity, we can compare *PhenoNet* with other existing methods. Firstly, we tested the performance of *PhenoNet* with the functional similarity based on GO biological process category. In general, *PhenoNet* performs better than other methods (Table 1). Figure 7 shows that similarities based on all three methods are correlated with GO BP category-based similarities. Quantificationally, *Pearson* correlation coefficient (*PCC*) scores [52] of *PhenoNet* on median and mean value are both 0.96, which are higher than the second best method Menche’s (0.82 and 0.89). Furthermore, the  $R^2$  scores [53] of *PhenoNet* on the median and mean value are 0.93 and 0.92, which are 0.24 and 0.13 higher than the second best method Menche’s respectively. Similarly, we also compared the



**Fig. 8** Phenotype similarity versus GO term similarity. The x-axis shows the similarity calculated by three methods: ontology-based (a), Menche’s method (b) and *PhenoNet* (c). The y-axis is GO term similarity based on molecular function category



three methods based on the GO molecular function (MF) category. Figure 8 shows that only *PhenoNet* similarities increase steadily with the increase of the GO MF-based similarities. In general, the performance of *PhenoNet* is significantly higher than other methods. Specifically, *Pearson* correlation coefficient (*PCC*) scores of *PhenoNet* on the median and mean value are 0.94 and 0.95 respectively, which are 0.33 and 0.29 higher than the second best method Menche's. In addition, the  $R^2$  scores of *PhenoNet* on median and mean value are more than two times of the second best method Menche's, which are 0.37 and 0.44 respectively.

All the results indicate that *PhenoNet* can calculate the similarity between phenotypes effectively and has stronger correlation with GO term-based similarity than other existing methods.

**Performance evaluation on co-expression dataset**

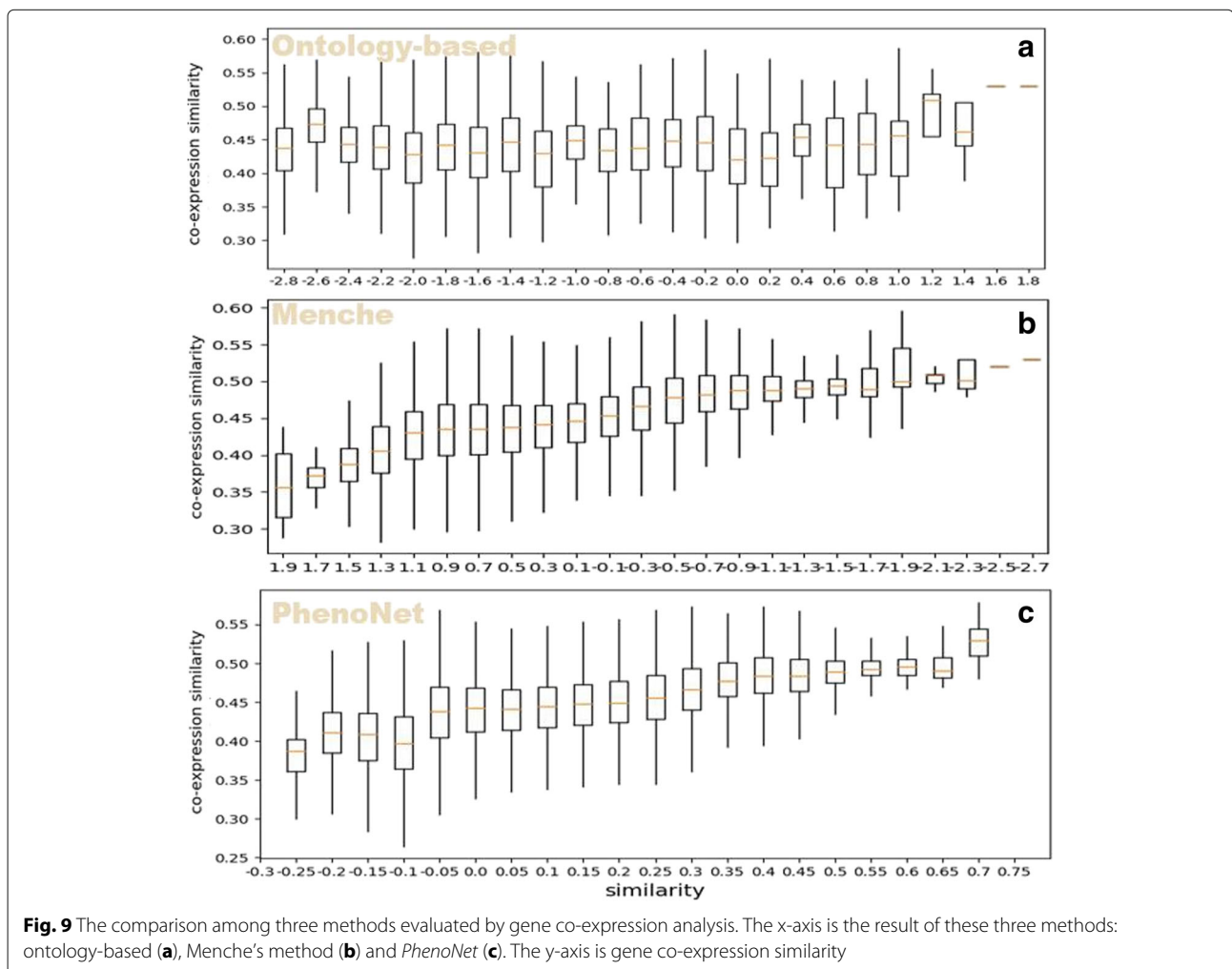
To evaluate the performance of *PhenoNet*, we also used the expression similarity of genes associated with two

phenotypes as an independent evidence to test the compared three methods. This evaluation method was borrowed from previous research [20].

We first computed the *Spearman* correlation coefficient  $\rho(g_i, g_j)$  [54] between gene  $g_i$  and gene  $g_j$ . Given two phenotypes  $p_1$  and  $p_2$ , let  $G_1$  and  $G_2$  be the protein sets associated with  $p_1$  and  $p_2$  respectively. The expression similarity between two phenotypes is defined as the average of  $|\rho(g_i, g_j)|$  over all protein pairs between  $G_1$  and  $G_2$  (Eq. 11).

$$Sim_{phe}(p1, p2) = \frac{1}{|G_1| \times |G_2|} \sum_{g_i \in G_1, g_j \in G_2} |\rho(g_i, g_j)| \tag{11}$$

We also compared our method with ontology-based and Menche's method. The result shows that both *PhenoNet* and Menche's method are correlated with the expression similarity (Fig. 9). There is no correlation trend between ontology-based method and expression similarity. In



**Fig. 9** The comparison among three methods evaluated by gene co-expression analysis. The x-axis is the result of these three methods: ontology-based (a), Menche's method (b) and *PhenoNet* (c). The y-axis is gene co-expression similarity

**Table 2** The Pearson correlation coefficient and  $R^2$  score between three methods and gene co-expression similarity

Measure		Ontology-based	Menche	PhenoNet
PCC	Median	0.42	0.93	<b>0.96</b>
	Mean	0.42	0.94	<b>0.97</b>
$R^2$	Median	0.16	0.86	<b>0.93</b>
	Mean	0.17	0.90	<b>0.95</b>

The highest values were highlighted in boldface

general, *PhenoNet* performs better than other methods. We also calculated *Pearson* correlation coefficient (*PCC*) scores of every method on the median and mean value of scores in each bar in Fig. 9. *Pearson* correlation coefficient (*PCC*) score of *PhenoNet* on median is 0.96, while the scores of Menche’s method and ontology-based method are 0.93 and 0.42 respectively. *Pearson* correlation coefficient (*PCC*) score of *PhenoNet* on mean is 0.97, while the scores of Menche’s method and ontology-based method are 0.94 and 0.42 respectively (Table 2). In addition, we also calculated the  $R^2$  score for each method. The  $R^2$  scores of *PhenoNet* on median and mean value are 0.93 and 0.95, which are higher than the second best Menche’s method (0.86 and 0.90 respectively).

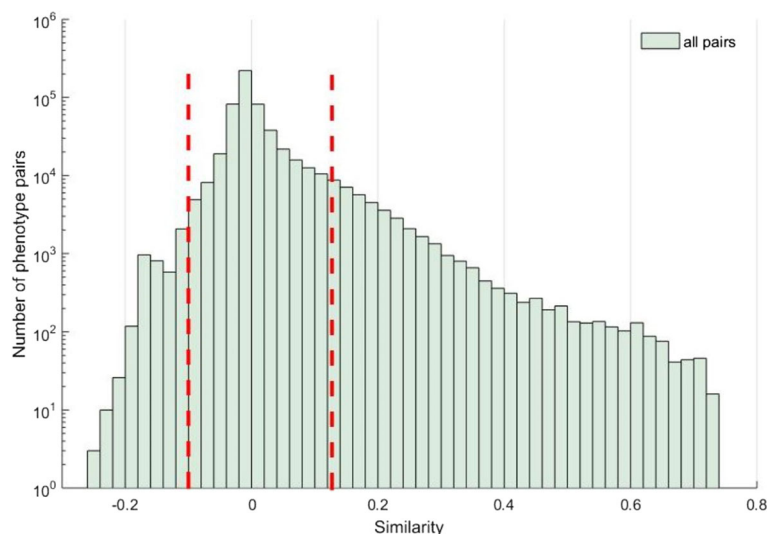
These results indicate that *PhenoNet* can calculate the similarity between phenotypes effectively and has stronger correlation with expression similarity than other existing methods.

### Significance of differentiating similar and dissimilar phenotypes

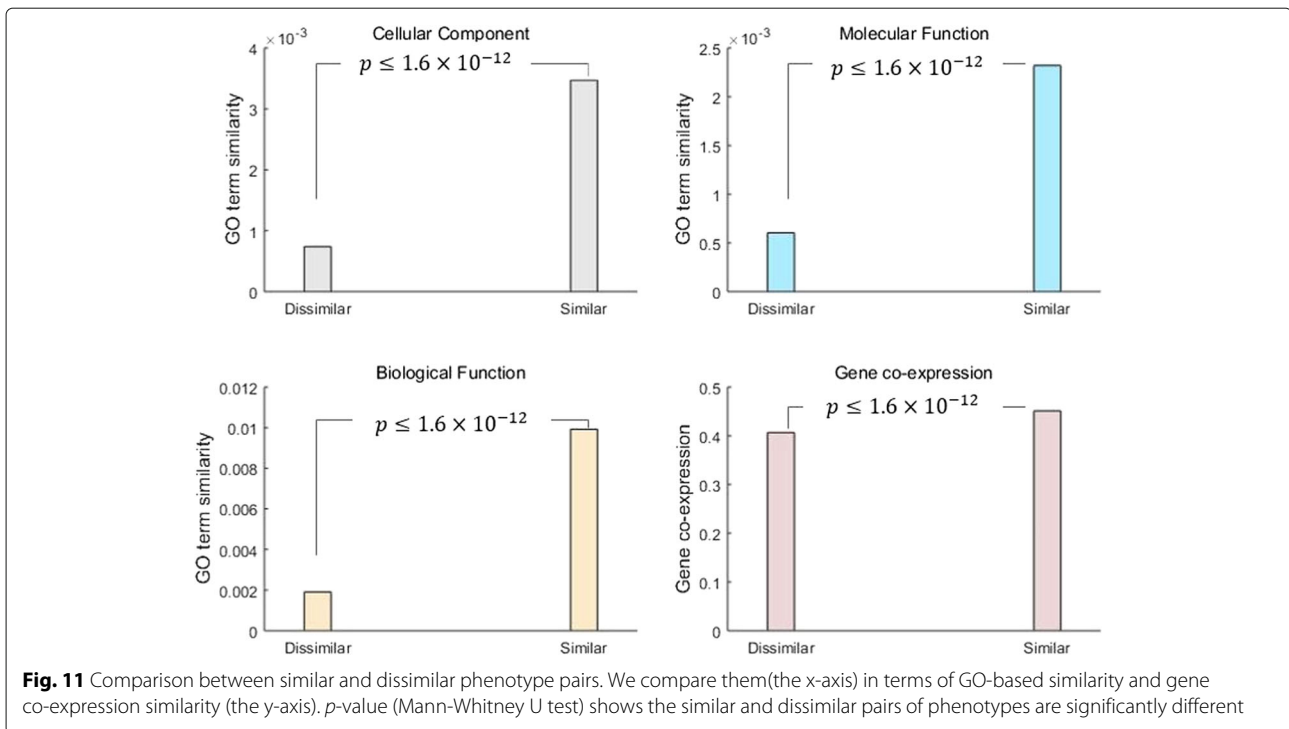
To test whether *PhenoNet* can differentiate similar and dissimilar phenotypes efficiently, we compared the

*PhenoNet* score of similar phenotype pairs and dissimilar phenotype pairs. The phenotype pairs with “high” and “low” *PhenoNet* scores are defined as similar and dissimilar set respectively. Specifically, the similar and dissimilar phenotype pairs are defined based on the following statistical model. Firstly, 100,000 pairs of phenotypes are selected randomly, saved as  $P_{ran}$ . The similarities of phenotype pairs in  $P_{ran}$  are calculated to obtain the distribution of the similarities. Secondly, we computed the  $p$ -value of each pair. Phenotype pairs with  $p$ -value  $\leq -0.05$  and  $p$ -value  $\geq 0.05$  are defined as dissimilar and similar phenotype pairs respectively. Figure 10 shows the distribution of similarities. The two red lines in Fig. 10 represent the boundary of similarity and dissimilarity.

After we obtain the similar and dissimilar set, we can test whether the similar and dissimilar pairs of phenotypes are significantly different based on different independent evidences. The similarities of phenotype pairs are calculated based on three GO categories (MF, CC, BP) and gene expression data respectively. Then, we test whether the similarities of similar set and dissimilar set are significantly different. The results show that the similarities of similar set and dissimilar set are significantly different (Fig. 11, Mann-Whitney U test [55],  $p$ -value  $\leq 1.6 \times 10^{-12}$ ). In Fig. 11, the bars represent the mean value of the similarities of phenotype pairs in the similar or dissimilar set. The results show that the mean values of dissimilar set are smaller than the mean values of similar set based on all four independence evidences. These results indicate that *PhenoNet* can differentiate similar and dissimilar phenotypes efficiently.



**Fig. 10** Similarity distribution of all phenotype pairs. The similarity ranges from -0.25 to 0.75, and the red lines indicate the boundary similarity scores of significantly similar or dissimilar pairs



## Conclusions

Phenotype similarity calculation plays a key role in disease diagnosis. Recently, some methods have been proposed to measure the phenotype similarity. These measures can be grouped into three categories: text mining-based, ontology-based. However, the existing methods can not distinguish the similarities between terms with the same common ancestor and ignore the interactions between annotated proteins by phenotypes. In this paper, we proposed a network based method, called *PhenoNet*, to calculate the phenotype similarity. *PhenoNet* includes three steps: network modules identification, network localization, phenotype internal similarity calculation and phenotype similarity calculation. The network localization test shows that phenotypes can be represented by network modules. We compared *PhenoNet* with two existing methods: ontology-based method and Menche's method. Furthermore, based on two independent evaluation datasets (gene ontology and gene co-expression data), evaluation test shows that *PhenoNet* performs better than existing methods. Our work opens a new window for phenotype similarity calculation, which may be potentially helpful to disease diagnosis.

## Acknowledgements

This work was supported by National Natural Science Foundation of China (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. 61602386 and 61332014).

## Funding

The publication costs for this article were funded by the corresponding author's institution.

## Availability of data and materials

The datasets during and/or analysed during the current study available from the corresponding author on reasonable request.

## About this supplement

This article has been published as part of *BMC Bioinformatics* Volume 19 Supplement 5, 2018: Selected articles from the Biological Ontologies and Knowledge bases workshop 2017. The full contents of the supplement are available online at <https://bmcbioinformatics.biomedcentral.com/articles/supplements/volume-19-supplement-5>.

## Authors' contributions

JP and XS designed the web tool framework; WH implemented the web tool; JP and WH wrote this manuscript. 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

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Published: 11 April 2018

## References

1. De Ligjt J, Willemsen MH, Van Bon BW, Kleefstra T, Yntema HG, Kroes T, Vulto-van Silfhout AT, Koolen DA, De Vries P, Gilissen C, et al. Diagnostic

- exome sequencing in persons with severe intellectual disability. *N Engl J Med.* 2012;367(20):1921–9.
2. Cheng L, Jiang Y, Wang Z, Shi H, Sun J, Yang H, Zhang S, Hu Y, Zhou M. Dissim: an online system for exploring significant similar diseases and exhibiting potential therapeutic drugs. *Sci Rep.* 2016;6:30024.
  3. Hu Y, Zhou M, Shi H, Ju H, Jiang Q, Cheng L. Measuring disease similarity and predicting disease-related ncRNAs by a novel method. *BMC Med Genomics.* 2017;10(5):71. <https://doi.org/10.1186/s12920-017-0315-9>.
  4. Hu Y, Zhao L, Liu Z, Ju H, Shi H, Xu P, Wang Y, Cheng L. Dissim: an online system for calculating similarity between disease sets. *J Biomed Semant.* 2017;8(1):28.
  5. Zemojtel T, Köhler S, Mackenroth L, Jäger M, Hecht J, Krawitz P, Graul-Neumann L, Doelken S, Ehme K, Spielmann M, et al. Effective diagnosis of genetic disease by computational phenotype analysis of the disease-associated genome. *Sci Transl Med.* 2014;6(252):252–123252123.
  6. Robinson PN. Deep phenotyping for precision medicine. *Hum Mutat.* 2012;33(5):777.
  7. Van Driel MA, Bruggeman J, Vriend G, Brunner HG, Leunissen JA. A text-mining analysis of the human phenome. *Eur J Hum Genet.* 2006;14(5):535–42.
  8. Botstein D, Risch N. Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease. *Nat Genet.* 2003;33:228–37.
  9. Oti M, Brunner HG. The modular nature of genetic diseases. *Clin Genet.* 2007;71(1):1–11.
  10. Liu G, Jiang Q. Alzheimer's disease cd33 rs3865444 variant does not contribute to cognitive performance. *Proc Natl Acad Sci.* 2016;113(12):1589–90.
  11. Mathur S, Dinakarpanand D. Finding disease similarity based on implicit semantic similarity. *J Biomed Inform.* 2012;45(2):363–71.
  12. Deans AR, Lewis SE, Huala E, Anzaldo SS, Ashburner M, Balhoff JP, Blackburn DC, Blake JA, Burleigh JG, Chanet B, et al. Finding our way through phenotypes. *PLoS Biol.* 2015;13(1):1002033.
  13. 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.
  14. Yang H, Robinson PN, Wang K. Phenolyzer: phenotype-based prioritization of candidate genes for human diseases. *Nat Methods.* 2015;12(9):841–3.
  15. Freimer N, Sabatti C. The human phenome project. *Nat Genet.* 2003;34(1):15.
  16. Jiang L, Gong B, Xi C, Tao L, Chao W, Fan Z, Li C, Xiang L, Rao S, Xia L. Dosim: An R package for similarity between diseases based on disease ontology. *Bmc Bioinformatics.* 2011;12(1):266.
  17. Batet M, Sánchez D, Valls A. An ontology-based measure to compute semantic similarity in biomedicine. *J Biomed Inform.* 2011;44(1):118–25.
  18. Ji X, Ritter A, Yen PY. Using ontology-based semantic similarity to facilitate the article screening process for systematic reviews. *J Biomed Inform.* 2017;69:33–42.
  19. Jiang R, Gan M, He P. Constructing a gene semantic similarity network for the inference of disease genes. *BMC Syst Biol.* 2011;5(2):1–11.
  20. Menche J, Sharma A, Kitsak M, Ghiassian SD, Vidal M, Loscalzo J, Barabási AL. Disease networks: uncovering disease-disease relationships through the incomplete interactome. *Science.* 2015;347(6224):1257601.
  21. Groza T, Kohler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, Schriml LM, Kibbe WA, Schofield PN, Beck T, et al. The human phenotype ontology: Semantic unification of common and rare disease. *Am J Hum Genet.* 2015;97(1):111–24.
  22. Le D, Dang V. Ontology-based disease similarity network for disease gene prediction. *Vietnam J Comput Sci.* 2016;3(3):197–205.
  23. 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.
  24. Liang C, Jie S, Xu W, Dong L, Yang H, Meng Z. Oahg: an integrated resource for annotating human genes with multi-level ontologies. *Sci Rep.* 2016;6:34820.
  25. Hao J, Sun J, Chen G, Wang Z, Yu C, Ming Z. Efficient and robust emergence of norms through heuristic collective learning. *ACM Trans Auton Adapt Syst (TAAS).* 2017;12(4):23.
  26. Hao J, Huang D, Cai Y, Leung H-f. The dynamics of reinforcement social learning in networked cooperative multiagent systems. *Eng Appl Artif Intell.* 2017;58:111–22.
  27. Köhler S, Schulz MH, Krawitz P, Bauer S, Dölken S, Ott CE, Mundlos C, Horn D, Mundlos S, Robinson PN. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *Am J Hum Genet.* 2009;85(4):457–64.
  28. Masino AJ, Dechene ET, Dulik MC, Wilkens A, Spinner NB, Krantz ID, Pennington JW, Robinson PN, White PS. Clinical phenotype-based gene prioritization: an initial study using semantic similarity and the human phenotype ontology. *BMC Bioinformatics.* 2014;15(1):248.
  29. Robinson PN, Köhler S, Bauer S, Seelow D, Horn D, Mundlos S. The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am J Hum Genet.* 2008;83(5):610–5.
  30. Kahanda I, Funk C, Verspoor K, Ben-Hur A. Phenostruct: Prediction of human phenotype ontology terms using heterogeneous data sources. *F1000research.* 2015;4:259.
  31. Deng Y, Gao L, Wang B, Guo X. Hposim: An R package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology. *Plos ONE.* 2015;10(2):0115692.
  32. Westbury SK, Turro E, Greene D, Lentaigne C, Kelly AM, Bariana TK, Simeoni I, Pillois X, Attwood A, Austin S. Human phenotype ontology annotation and cluster analysis to unravel genetic defects in 707 cases with unexplained bleeding and platelet disorders. *Genome Med.* 2015;7(1):36.
  33. 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. <https://doi.org/10.1186/s12859-017-1959-3>.
  34. Peng J, Lu J, Shang X, Chen J. Identifying consistent disease subnetworks using dnet. *Methods.* 2017;131:104–10.
  35. Peng J, Zhang X, Hui W, Lu J, Li Q, Liu S, Shang X. Improving the measurement of semantic similarity by combining gene ontology and co-functional network: a random walk based approach. *BMC Syst Biol.* 2018;12(suppl 12). in press.
  36. Hu J, Shang X. Detection of network motif based on a novel graph canonization algorithm from transcriptional regulation networks. *Molecules.* 2017;22(12):2194.
  37. Berriz GF, Beaver JE, Cenik C, Tasan M, Roth FP. Next generation software for functional trend analysis. *Bioinformatics.* 2009;25(22):3043–44.
  38. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al. *Proc Natl Acad Sci U S A.* 2004;101(16):6062–7.
  39. Matys V, Fricke E, Geffers R, Gössling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV. Transfac register: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 2003;31(1):374–8.
  40. Rolland T, Tasan M, Charlotiaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, et al. A proteome-scale map of the human interactome network. *Cell.* 2014;159(5):1212–26.
  41. Venkatesan K, Rual J, Vazquez A, Stelz U, Lemmens I, Hirozaneekishikawa T, Hao T, Zenkner M, Xin X, Goh K, et al. An empirical framework for binary interactome mapping. *Nat Methods.* 2009;6(1):83–90.
  42. Stelz U, Worm U, Lalowski M, Haenic G, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S. A human protein-protein interaction network: a resource for annotating the proteome. *Cell.* 2005;122(6):957.
  43. Rual JF, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz GF, Gibbons FD, Dreze M, Ayivi-Guedehoussou N. Towards a proteome-scale map of the human protein-protein interaction network. *Nature.* 2005;437(7062):1173–8.
  44. Yu H, Leah T, Stanley T, Evan W, Fana G, Fan C, Nenad S, Tomoko HK, Edward R, Yang X. Leveraging the power of next-generation sequencing to generate interactome datasets. *Nat Methods.* 2011;8(6):478.
  45. Licata L, Briganti L, Peluso D, Perfetto L, Iannuccelli M, Galeota E, Sacco F, Palma A, Nardoza AP, Santonicio E. Mint, the molecular interaction database: 2012 update. *Nucleic Acids Res.* 2007;35(Database issue):572–4.
  46. Stark C, Breitkreutz BJ, Chatranyamontri A, Boucher L, Oughtred R, Livstone MS, Nixon J, Van AK, Wang X, Shi X. The biogrid interaction database: 2011 update. *Nucleic Acids Res.* 2015;43(Database issue):470.

47. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A. Human protein reference database—2009 update. *Nucleic Acids Res.* 2009;37(Database issue):767.
48. Lee DS, Park J, Kay KA, Christakis NA, Oltvai ZN, Barabási AL. The implications of human metabolic network topology for disease comorbidity. *Proc Natl Acad Sci U S A.* 2008;105(29):9880.
49. Ruepp A, Brauner B, Dungerkaltenbach I, Frishman G, Montrone C, Stransky M, Waegel B, Schmidt T, Doudieu ON, Stümpflen V. Corum: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.* 2010;38(Database issue):497.
50. Hornbeck PV, Kornhauser JM, Tkachev S, Zhang B, Skrzypek E, Murray B, Latham V, Sullivan M. Phosphositeplus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* 2012;40(Database issue):261.
51. Vinayagam A, Stelzl U, Foulle R, Plassmann S, Zenkner M, Timm J, Assmus HE, Andrade-Navarro MA, Wanker EE. A directed protein interaction network for investigating intracellular signal transduction. *Sci Signal.* 2011;4(189):8.
52. Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. In: *Noise Reduction in Speech Processing.* Berlin: Springer Berlin Heidelberg; 2009. p. 1–4.
53. Lewis-Beck MS. "R-squared" Thousand Oaks, Calif. The Sage Encyclopedia of Social Science Research Methods. 2004. [http://works.bepress.com/michael\\_lewis\\_beck/126/](http://works.bepress.com/michael_lewis_beck/126/).
54. Myers L, Siros MJ. Spearman Correlation Coefficients, Differences between. In: *Wiley StatsRef: Statistics Reference Online.* Wiley; 2014. <https://doi.org/10.1002/9781118445112.stat02802>.
55. McKnight PE, Najab J. Mann-Whitney U Test. In: *The Corsini Encyclopedia of Psychology.* Wiley; 2010. <https://doi.org/10.1002/9780470479216.corpsy0524>.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at  
[www.biomedcentral.com/submit](http://www.biomedcentral.com/submit)

