Analyzing factors involved in the HPO-based semantic similarity calculation

Jiajie Peng*, Qianqian Li*, Bolin Chen*, Jialu Hu* and Xuequn Shang†
*School of Computer Science
Northwestern Polytechnical University, Xi'an, China
Email: jiajiepeng@nwpu.edu.cn
†School of Computer Science
Northwestern Polytechnical University, Xi'an, China
Email: shang@nwpu.edu.cn, Corresponding author

Abstract—**Although disease diagnosis have greatly benefited from next generation sequencing technologies, it is still difficult to make the right diagnosis based on purely sequencing technologies for many diseases with complex phenotypes and high genetic heterogeneity. Recently, calculating Human Phenotype Ontology (HPO)-based phenotype semantic similarity has contributed a lot for completing disease diagnosis. However, factors which affect the accuracy of HPO-based semantic similarity have not been evaluated systematically. In this study, we propose a new framework called $HPOFactor$ to evaluate these factors.**

I. Introduction

In the last few years, disease diagnosis has greatly benefited from the rapid development of next generation sequencing (NGS) technologies [1], [2], [3]. However, it is difficult to make the right diagnosis based on purely sequencing technologies for many diseases with complex phenotypes and high genetic heterogeneity.

Many cases show that ontology is effective to represent biomedical information as terms and their directed relationships with a directed acyclic graph (DAG) [4], [5], [6], [7]. In order to meet the demand, an ontology called Human Phenotype Ontology (HPO) was constructed to describe the abnormal human phenotypes encountered in human Mendelian disease by Robinson *et al* in 2008 [8].

Although ontology-based semantic similarity measurement has been extensively studied in the last ten years [9], [10], [11], [12], [13], it is still a difficult task to measure the phenotype similarity based on HPO structure and annotations. The reason is that many factors could affect the accuracy of HPO-based phenotype semantic similarity. [14].

To figure out how different factors affect the performance of ontology-based semantic similarity measurement, some methods have been proposed to evaluate different involved factors. To test whether different editions of Gene Ontology (GO) would result in different semantic similarities, Gillis *et al* proposed an evaluation framework based on protein interaction networks [15]. Skunca *et al* proposed a novel method to systematically evaluate the quality of the computationally inferred GO annotations [16]. Both of the aforementioned methods are based on the historical versions of ontology. These methods cannot be used to evaluate the factors that affect the performance of HPO-based semantic similarity measurement, since the historical versions of HPO are not available currently (personal communication with the founder of HPO). Furthermore, other factors may also affect the accuracy of HPO-based semantic similarity.

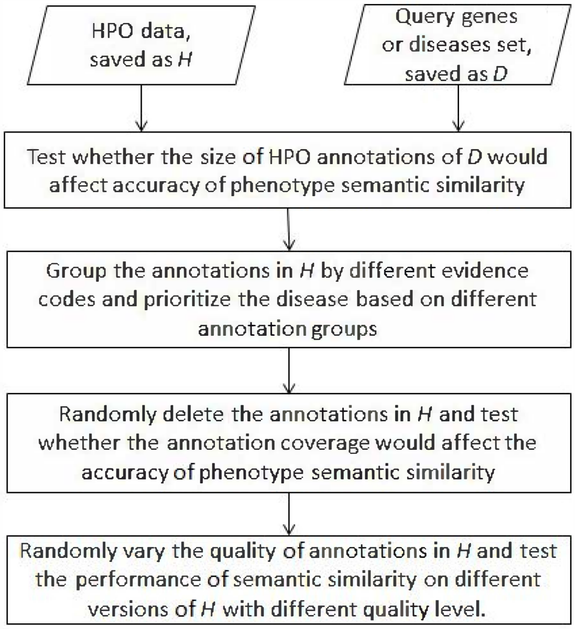
HPO data, saved as *H*
Query genes or diseases set, saved as *D*
Test whether the size of HPO annotations of *D* would affect accuracy of phenotype semantic similarity
Group the annotations in *H* by different evidence codes and prioritize the disease based on different annotation groups
Randomly delete the annotations in *H* and test whether the annotation coverage would affect the accuracy of phenotype semantic similarity
Randomly vary the quality of annotations in *H* and test the performance of semantic similarity on different versions of *H* with different quality level.

Fig. 1: The workflow of HPOFactor.

II. Methods

We proposed $HPOFactor$, a new framework to evaluate the factors that affect the performance of phenotype semantic similarity measurement based on human phenotype ontology (HPO). The proposed framework has four parts. The diagram of the whole framework is shown in Figure 1.

A. Calculating HPO-based semantic similarity

HPO provides a structured and controlled vocabulary to describe the human phenotypes and the genes/diseases associated with the phenotypes [8]. Since the phenotype sets of patient, gene and disease are all able to be unified by HPO terms,

978-1-5090-1610-5/16/$31.00 ©2016 IEEE

calculating the similarity between patient and gene/disease is equal to calculating the similarity between two sets of HPO terms.

Let P_1 and P_2 be two phenotype term sets corresponding to a patient and a disease (or gene) respectively. Their HPO-based similarity is calculated as follows.

$$sim(P_1, P_2) = \frac{1}{2} \times sim_{set}(P_1 \rightarrow P_2) + \frac{1}{2} \times sim_{set}(P_2 \rightarrow P_1) \tag{1}$$

where $sim_{set}(P_1 \rightarrow P_2)$ represents the similarity from P_1 to P_2. Mathematically, $sim_{set}(P_1 \rightarrow P_2)$ is defined as follows.

$$sim_{set}(P_1 \rightarrow P_2) = avg\left[\sum_{p_1 \in P_1} max_{p_2 \in P_2} sim_{term}(p_1, p_2)\right] \tag{2}$$

where $sim_{term}(p_1, p_2)$ represents semantic similarity between two phenotypes p_1 and p_2. It is noted that the similarity from phenotype set P_1 to P_2 is different from the similarity from phenotype set P_2 to P_1. Therefore, equation 1 averages the two dissymmetric similarities as the similarity between two phenotype sets.

To calculate $sim_{term}(p_1, p_2)$, let $S(p_1, p_2)$ be the set of all common ancestors of p_1 and p_2. p_{min} is the term that has the minimal annotations in $S(p_1, p_2)$. Given two phenotypes p_1 and p_2, their similarity $sim_{term}(p_1, p_2)$ is defined as follows.

$$sim_{term}(p_1, p_2) = -\log \frac{N_{p_{min}}}{N} \tag{3}$$

where $N_{p_{min}}$ is the number of annotations of p_{min} (including annotations of itself and its descendants) and N is the total number of annotations involved in HPO.

B. Test the effect of the size of annotation set

In this subsection, we proposed a method to test whether the size of annotation set would affect the precision of semantic similarity.

Given a set of query patients Q, each element q in Q has an annotation set obtained from clinical treatment saved as P_q. Given a set of genes/diseases H involved in HPO database, each element h in H has an annotation set obtained from HPO database saved as P_h. We changed the size threshold of annotations s and calculate the semantic similarity at different thresholds. Given the threshold s, the detail of the method is described as follows. For each element h in H, we randomly selected s phenotypes from P_h, saved as P_h^s. This step is represented mathematically in Equation 4.

$$P_h^s = RandomSelection(P_h, s) \tag{4}$$

For each query patient q in Q, we calculate the similarity between P_q and P_h^s for each h in H using Equation 1.At last, we can test whether the known patient associated element (gene or disease) has a high rank in H_{order}. The higher the rank is, the better the performance of the semantic similarity measurement is.

C. Test the effect of annotations with different evidence codes

In this subsection, we test whether using different annotations with different evidence codes would affect the precision of HPO-based semantic similarity. Given the annotation set A, A_e represents the annotation set with evidence e. Given a set of genes/diseases H, the annotation set of each element h in H is obtained from A_e, saved as P_{he}. Similar with the process described in last subsection, we rank all elements in H based on the similarities with the phenotypes of query patient.

$$H_{order} = Rank(H, \{y | y = sim(P_q, P_{he}), h \in H\}) \tag{5}$$

Finally, we could see which evidence code can lead the best performance.

D. Test the effect of annotation quality

To determine whether annotation quality was one of the factors that control the performance of HPO-based semantic similarity, we varied the HPO annotation quality by randomly swapping the phenotype-annotation associations in HPO. For example, assume that $d_1 \rightarrow p_1$ and $d_2 \rightarrow p_2$ are two disease-phenotype pairs randomly selected from HPO. After the swapping process, we get two new pairs $d_1 \rightarrow p_2$ and $d_2 \rightarrow p_1$ to replace the original two pairs. Given the original HPO annotation set A, we can generate a low quality set A_u by randomly swapping the phenotype-annotation associations. u represents different quality levels.A_u has the same size with A but different quality level. For each low quality level u, we use the low quality annotation set A_u to calculate the semantic similarity between phenotypes.

E. Test the effect of annotation coverage

Currently, HPO is not complete. Therefore, it is critical to test whether annotation coverage was a key factor for HPO-based semantic similarity measurement. To this end, we randomly delete the annotations from annotation set A to generate a low coverage annotation set A_c. c represents different coverage levels. For each coverage level c, we use the low coverage annotation set A_c to calculate the semantic similarity between phenotypes.

III. Results

A. Data preparation

The Human Phenotype Ontology (HPO) data used in our experiment was downloaded from the HPO official website (http://human-phenotype-ontology.github.io/) on April 1st, 2016. It includes $459,452$ gene annotations and $78,313$ disease annotations. $HPOFactor$ was implemented with Python language.

We used the curated clinical phenotype features in [14] to generate simulated patients for experiments. The associated phenotypes, disease causative genes and penetrance of each phenotype are available in the dataset. For each disease, we simulated 100 patients. To consider the gender-specificity of phenotypes, we first simulated the gender of each patient. A

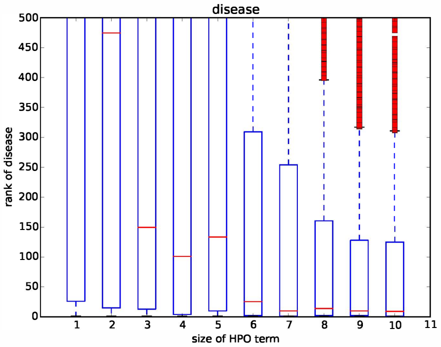

Fig. 2: Distribution of the rank of the patient associated disease by changing the size of phenotype annotation set. The x-axis is the number of HPO annotations. The y-axis is the rank of disease associated with the query patient.

random number f_g was generated. Then, the patient's gender is assigned as follows:

$$\begin{cases} f_g > 0.5 & \text{,male} \\ f_g \leq 0.5 & \text{,famale} \end{cases} \quad (6)$$

Second, given a phenotype p of a patient, a random number r_p was generated. Let f_p be the penetrance of this phenotype associated with the assigned disease. If $r_p < f_p$, the phenotype p was assigned to the patient. It is noted that each simulated patient must have at least one phenotype. Finally, 3300 patients was generated.

B. Evaluation for the size of annotation set

In this experiment, we compared the results of using different sizes of annotation set to identify the disease associated with the patient. Figure 2 shows that the performance improved with the increase of the size of annotation set. Noted that the performance become stable when $s > 5$.

C. Evaluation for the annotations with different evidence codes

In this part, we test whether using annotations with different evidence codes would affect result of identifying the disease associated with the patient. We do not test the performance for causative gene identification since the gene annotations in HPO do not have evidence codes currently. We only compare three evidence codes: IEA, TAS and PCS, since other evidence codes do not have enough number of annotations. To avoid the bias resulting from the lack of annotation, we did the experiment on the size of annotations sets which are not larger than 5. We choose the size threshold since the experiment in last subsection shows that the performance become stable when $s > 5$.

Figure 3 shows that annotations with PCS evidence code perform better the the annotations with IEA and TAS evidence code.

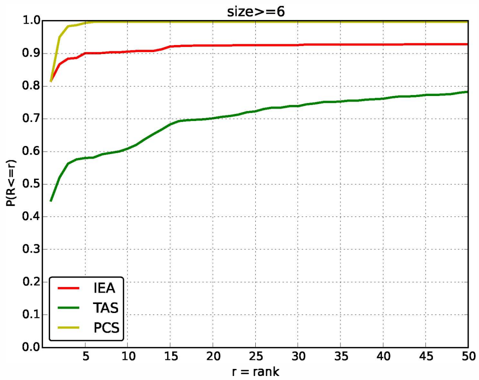

Fig. 3: Distribution of the rank of the disease by fixing the size of phenotype annotations with different evidence code. The x-axis is the ranking threshold for the disease. The y-axis is the ratio of patients satisfying the ranking threshold.

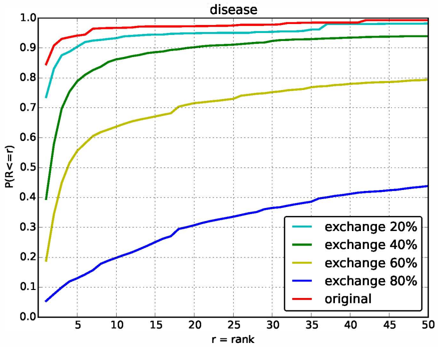

Fig. 4: Distribution of the rank of the disease by varying the quality of phenotype annotations. The x-axis is the ranking threshold for the disease/causative gene. The y-axis is the ratio of patients satisfying the ranking threshold.

D. Evaluation for the annotation quality

To test the effect of annotation quality to the performance of HPO-based semantic similarity, we compared the results of using annotation sets with different qualities to identify the patient associated diseases (Figure 4). Overall, the result shows that the performance goes down with the decrease of the annotation quality in both experiments. It is shows that the performance decreases significantly when more than 40% annotations become noise.

Furthermore, the statistical test shows that the result for original annotation set is significantly different with 40%, 60% and 80% set (Tukey test, p-value < 0.05).

E. Evaluation for the annotation coverage

To test the effect of annotation coverage to the performance of HPO-based semantic similarity, we randomly delete the annotations and use annotation sets with different coverage to identify the associated disease and causative genes. The result shows that the performance of HPO-based semantic similarity

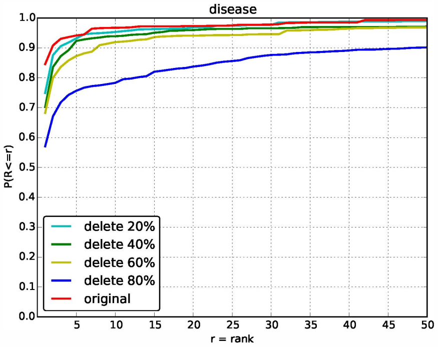

Fig. 5: Distribution of the rank of the disease by changing the coverage of phenotype annotations. The x-axis is the rank for the disease/causative gene. The y-axis is the ratio of patients satisfying the ranking threshold.

decreased with the reduction of the annotations (Figure 5). However, there was no significant difference when the deleted annotations are less than 60% (Tukey test, p-value > 0.05). It indicates that HPO-based semantic similarity is more sensitive to the quality of annotations than the coverage of annotations.

IV. Conclusion

In this article, we proposed a novel framework called $HPOFactor$ to evaluate the factors that may affect the accuracy of HPO-based semantic similarity. $HPOFactor$ evaluates four factors involved in the HPO-based semantic similarity: size of annotation set, evidence code of annotations, quality of annotations and coverage of annotations. The evaluation result can make the HPO-based semantic similarity better used in phenotype-based causative gene prediction and disease prediction. In future, we will design semantic similarity measurement based on the characteristic of these factors.

References

[1] J. De Ligt, M. H. Willemsen, B. W. van Bon, T. Kleefstra, H. G. Yntema, T. Kroes, A. T. Vulto-van Silfhout, D. A. Koolen, P. de Vries, C. Gilissen *et al.*, "Diagnostic exome sequencing in persons with severe intellectual disability," *New England Journal of Medicine*, vol. 367, no. 20, pp. 1921–1929, 2012.

[2] Y. Yang, D. M. Muzny, F. Xia, Z. Niu, R. Person, Y. Ding, P. Ward, A. Braxton, M. Wang, C. Buhay *et al.*, "Molecular findings among patients referred for clinical whole-exome sequencing," *Jama*, vol. 312, no. 18, pp. 1870–1879, 2014.

[3] T. D. D. D. Study, "Large-scale discovery of novel genetic causes of developmental disorders," *Nature*, vol. 519, no. 7542, pp. 223–228, 2015.

[4] J. Peng, T. Wang, J. Wang, Y. Wang, and J. Chen, "Extending gene ontology with gene association networks," *Bioinformatics*, vol. 32, no. 8, pp. 1185–1194, 2016.

[5] J. Dutkowski, M. Kramer, M. A. Surma, R. Balakrishnan, J. M. Cherry, N. J. Krogan, and T. Ideker, "A gene ontology inferred from molecular networks," *Nature biotechnology*, vol. 31, no. 1, pp. 38–45, 2013.

[6] Y. Hu, W. Zhou, J. Ren, L. Dong, Y. Wang, S. Jin, and L. Cheng, "Annotating the function of the human genome with gene ontology and disease ontology." *BioMed Research International*, 2016.

[7] W. X. L. D. Y. H. . M. Z. Liang Cheng, Jie Sun, "Oahg: an integrated resource for annotating human genes with multi-level ontologies," *Scientific reports*, vol. 10, p. 34820, 2016.

[8] P. N. Robinson, S. Köhler, S. Bauer, D. Seelow, D. Horn, and S. Mundlos, "The human phenotype ontology: a tool for annotating and analyzing human hereditary disease," *The American Journal of Human Genetics*, vol. 83, no. 5, pp. 610–615, 2008.

[9] J. Peng, H. Li, Q. Jiang, Y. Wang, and J. Chen, "An integrative approach for measuring semantic similarities using gene ontology," *BMC systems biology*, vol. 8, no. Suppl 5, p. S8, 2014.

[10] H. Caniza, A. E. Romero, S. Heron, H. Yang, A. Devoto, M. Frasca, M. Mesiti, G. Valentini, and A. Paccanaro, "Gossto: a stand-alone application and a web tool for calculating semantic similarities on the gene ontology," *Bioinformatics*, vol. 30, no. 15, pp. 2235–2236, 2014.

[11] J. Peng, Y. Wang, and J. Chen, "Towards integrative gene functional similarity measurement," *BMC bioinformatics*, vol. 15, no. 2, p. 1, 2014.

[12] J. Peng, H. Li, Y. Liu, L. Juan, Q. Jiang, Y. Wang, and C. Jin, "Intego2: a web tool for measuring and visualizing gene semantic similarities using gene ontology," *Bmc Genomics*, vol. 17, no. 5, 2016.

[13] J. Peng, S. Uygun, T. Kim, Y. Wang, S. Y. Rhee, and J. Chen, "Measuring semantic similarities by combining gene ontology annotations and gene co-function networks," *BMC bioinformatics*, vol. 16, no. 1, p. 1, 2015.

[14] A. J. Masino, E. T. Dechene, M. C. Dulik, A. Wilkens, N. B. Spinner, I. D. Krantz, J. W. Pennington, P. N. Robinson, and P. S. White, "Clinical phenotype-based gene prioritization: an initial study using semantic similarity and the human phenotype ontology," *BMC bioinformatics*, vol. 15, no. 1, p. 1, 2014.

[15] J. Gillis and P. Pavlidis, "Assessing identity, redundancy and confounds in gene ontology annotations over time," *Bioinformatics*, vol. 29, no. 4, pp. 476–82, 2013.

[16] N. Skunca, A. Altenhoff, and C. Dessimoz, "Quality of computationally inferred gene ontology annotations." *Plos Computational Biology*, vol. 8, no. 5, pp. e1 002 533–e1 002 533, 2012.