Measuring Phenotype Semantic Similarity using Human Phenotype Ontology

Jiajie Peng*, Hansheng Xue†, Yukai Shao†, Xuequn Shang‡ Yadong Wang§ and Jin Chen¶
*School of Computer Science
Northwestern Polytechnical University, Xi'an, China
Email: jiajiepeng@nwpu.edu.cn, Corresponding author
†School of Computer Science and Technology
Harbin Institute of Technology, Shenzhen, China
Email: xhs1892@gmail.com, shaoyk@bankcomm.com
‡School of Computer Science
Northwestern Polytechnical University, Xi'an, China
Email: shang@nwpu.edu.cn
§School of Computer Science and Technology
Harbin Institute of Technology, Harbin, China
Email: ydwang@hit.edu.cn, Corresponding author
¶Institute of Biomedical Informatics, College of Medicine
University of Kentucky, Lexington, KY 40536, USA
Email: chen.jin@uky.edu, Corresponding author

***Abstract*—It is critical yet remains to be challenging to make right disease diagnosis based on complex clinical characteristic and heterogeneous genetic background. Recently, Human Phenotype Ontology (HPO)-based phenotype similarity has been widely used to aid disease diagnosis. However, the existing measurements are revised based on the Gene Ontology-based term similarity models, which are not optimized for human phenotype ontologies. We propose a new similarity measure called $PhenoSim$. Our model includes a noise reduction component to model the noisy patient phenotype data, and a path-constrained Information Content-based method for measuring phenotype semantics similarity. Evaluation tests showed that $PhenoSim$ could improve the performance of HPO-based phenotype similarity measurement.**

I. Introduction

Patient phenotypes are usually defined as the observable characteristics of patients above the molecular level, such as anatomy, behavior, and biomedical properties [1]. Approaches that bridge the genetic variances and biological process activities with advanced phenotype data analysis have played a central role in deciphering gene or pathway functions in life science research [2], [3], [4]. A key step in those tools is to precisely measure phenotypic features, and combine such information into the framework of clinical disease diagnosis to improve clinical diagnosis efficiency. Therefore, a structured and controlled vocabulary, such as ontology, is often required.

In 2008, Robinson *et al* constructed an ontology namely Human Phenotype Ontology (HPO) to describe human phenotypic abnormalities that have been encountered in human disease [1]. Currently, HPO is widely used and usually combined with NGS data to improve disease diagnosis efficiency [5].

To improve diagnostic efficiency, computational tools have been developed to quantify the phenotypic similarity between patient symptoms and curated historical disease data or known phenotypes related with a gene [6], [7], [8]. Among them, computing HPO-based phenotype similarity plays a critical role in completing disease diagnosis process.

In literature, tools such as Phenomizer[6], OWLSim [9] and HPOSim [8] have been developed to exploit HPO-based semantic similarity. Several of them borrow ideas from Gene Ontology (GO)-based semantic similarity approaches, which have been extensively studied and widely used in the last decade [10], [11], [12].

Phenomizer and Masino *et al.* utilized information content (IC) to calculate the HPO-based semantic similarity between two phenotype ontology terms [6], [7]. PhenomeNet [13] and OWLSim [9] employ simGIC [14] to calculate the similarity between two sets of phenotype terms.

Although the aforementioned approaches have been widely used in clinical research, they calculate phenotype semantic similarities based on the methods optimized for measuring GO-based semantic similarity without taking the unique properties of HPO into account.

In this article, we present a new approach called $PhenoSim$. Comparing with the existing approaches, $PhenoSim$ has the following advantages:

- $PhenoSim$ is the first semantic similarity approach that is specially optimized for HPO;
- We develop a novel path-constrained Information Content (IC) to calculate the similarity between two HPO terms;
- $PhenoSim$ constructs a phenotype network and exploits a PageRank-based method to model the noises in the patient phenotype data set.

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

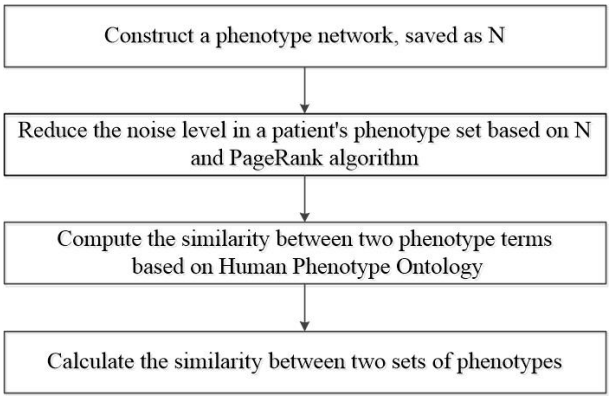

Fig. 1: The workflow of PhenoSim.

II. Methods

We propose $PhenoSim$, a new phenotype semantic similarity measurement optimized for Human Phenotype Ontologies (specifically HPO). $PhenoSim$ has four steps as shown in Figure 1.

A. Phenotype network construction

HPO provides structured and controlled vocabularies to describe human phenotypes [1]. It is generally understood that the phenotype terms associated with the same genes may be closely related to each other at the molecular level [15]. Hence, we identify the relationships between HPO terms using the genes associated with them.

We construct a phenotype network $N(V, E)$ based on the pair-wise Jaccard Index score between gene sets of two phenotypes. In $N(V, E)$, nodes in V represent phenotype terms. Two nodes are directly connected if the association score between them is larger than a user-given threshold (in our experiments, 0). The edge weight is the association score between two phenotype-associated gene sets..

B. Phenotype network noise reduction

It is technically challenging to precisely recognize all the patient phenotypes at the data collection step. Therefore, the noises in the patient phenotype data of HPO cannot be simply ignored [7]. To this end, we develop a new approach to reducing the noise level in patient's phenotype set P.

Given a patient's phenotype term set T, subnetwork of $N(V, E)$ called $N_T(T, E')$ can be generated using the approach in the previous subsection ($E' \subset E$, $T \subset V$). For a given disease, its corresponding correctly recognized phenotype terms are high similar to each other, in that their associated gene groups are highly overlapped [15]. Thus, we assume that in N_T the correctly recognized phenotype terms are the *important* nodes, and the associations between them are *high*. Based on the assumption, we differentiate the correctly recognized phenotype terms of a patient from noises using network topological properties such as node centrality, an index to describe the node significance in a network [16].

All the phenotypes in a patient's phenotype term set T are ranked based on their centrality. The top k phenotypes with the highest probabilities are selected as the well recognized phenotypes of the patient, denoted as T^k.

C. Measuring phenotype similarity

Sibling terms in the HPO structure are not necessary to have strong associations at gene level. Alternatively, semantically similar HPO terms are often "reachable", i.e., if two HPO terms t_1 and t_2 are similar, then there highly likely exists a directed path from one term to the other in the directed acyclic graph of HPO. Therefore, we define a new HPO term semantic similarity measurement as:

$$sim(t_1, t_2) = \begin{cases} \min\left(IC(t_1), IC(t_2)\right) & reachable \\ 0 & otherwise \end{cases} \quad (1)$$

where $IC(t)$ is the information content of phenotype term t, defined as $IC(t) = \ln \frac{U}{U_t}$, where U and U_t are the number of annotations associated with the root term and t, respectively (the annotations associated with all their descendants are also included). If t_1 and t_2 are reachable, the similarity is the minimum of their information contents. If t_1 and t_2 are unreachable, the similarity is 0.

D. Calculating phenotype set similarity

It is often required to predict whether a patient has certain disease or disease related gene. To this end, it is necessary to compare the phenotype set of a patient to all the phenotypes associated with a disease or a gene. While the patient phenotype set can be obtained in clinical treatment, the latter one are available in public databases such as OMIM [17].

Given a patient p_1 and a gene (or disease), let T_1^k and T_2^k be their associated phenotype term sets. T_1^k is the result of the noise reduction process in the previous subsection. T_2^k is set of phenotypes corresponding to the gene (or disease) obtained from the HPO database. We calculate the semantic similarity between the two patients based on the aggregation of the pair-wise similarities between terms across T_1^k and T_2^k by adopting the measure in [7].

$$Sim_{set}(T_1^k \to T_2^k) = \frac{1}{N_1} \sum_{t_i \in T_1^k} \max_{t_j \in T_2^k} sim(t_i, t_j) \quad (2)$$

$$Sim_{set}(T_2^k \to T_1^k) = \frac{1}{N_2} \sum_{t_j \in T_2^k} \max_{t_i \in T_1^k} sim(t_i, t_j) \quad (3)$$

where $sim(t_i, t_j)$ is the phenotype similarity calculated using Equation 1. N_1 and N_2 are the size of phenotype set T_1 and T_2 respectively. Note that since Equation 2 and 3 are asymmetric, the output depends on the order of the input. To avoid the asymmetry result, the similarity of two phenotype sets are calculated as:

$$Sim_{sym}(T_1^k, T_2^k) = \frac{1}{2}\left(Sim_{set}(T_1^k \to T_2^k) + Sim_{set}(T_2^k \to T_1^k)\right) \quad (4)$$

III. Results

A. Data preparation

The Human Phenotype Ontology (HPO) data were downloaded from the HPO official website (http://human-phenotype-ontology.github.io/) on July 4th, 2014. It includes $61,784$ phenotype-gene relationships and $99,186$ phenotype-disease relationships. $PhenoSim$ was implemented with Java SDK 7 and the JUNG library [18].

For performance evaluation, we first generated simulated patients based on the curated disease phenotype feature set used in [7]. In this dataset, for each of the 33 selected diseases, its disease causative genes, associated phenotypes, and penetrance of each phenotype are available. The patient simulation process is as follows. First, we randomly assign a disease to each patient. Second, for a given patient, we generated a random number between 0 and 1 (followed standard uniform distribution) for every phenotype associated with the assigned disease. If the random number was smaller than the penetrance value of the phenotype, the phenotype was assigned to the patient. Each simulated patient must have at least one phenotype. This set is named as the optimal phenotype set. We repeated the process for 100 times. As a result, $3,300$ simulated patients called "optimal patients with known causative genes", were generated.

The second evaluate data set is a simulation of the real clinical data. In the real clinical practice, patient phenotype sets often contain noise. Therefore, based on the optimal set, we generated a simulated patient set with added noise. Specifically, for every disease d, we randomly generated a large set of noise phenotype terms with the criterion that they (and their descendants) do not associate with any of the causative genes associated with d. For a given patient with disease d, we randomly selected noise phenotype terms of d and added them to the patient phenotype term set T, such that the number of noise terms is half of the optimal terms in the first dataset. Particularly, if a patient only had one optimal phenotype term, no noise term was added. Finally, $3,300$ simulated patients with noisy phenotype terms, called "noisy patient data with known causative genes", were simulated. In the dataset, for each simulated patient, there are in average 7.74 phenotype terms. The phenotype terms distribution is shown in Additional file 1.

Third, with the similar method, we generated patient sets with known diseases using data from OMIM [17], named "noisy patient data with known diseases".

B. Performance evaluation on causative gene prediction

We adopted the evaluation criterion from [7] to test whether the causative genes of a patient can be computationally identified. In this experiment, T_1^k and T_2^k are the phenotype sets corresponding to a simulated patient and a gene respectively. For a given patient, we computed the similarity score between every gene and the patient using $PhenoSim$, and then rank all the genes by their similarity scores from the largest to the smallest. If the causative gene's rank is higher than any other genesis, we conclude that $PhenoSim$ can accurately predict the causative gene. Similarly, we test the performance of four existing approaches, i.e. Masino [7], Lin [19], Jiang [20], and Schlicker [21], on the datasets described above.

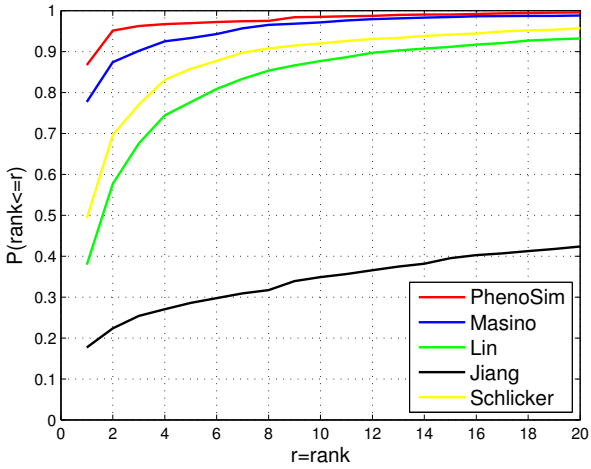

Fig. 2: Cumulative distribution of the rank of the causative genes on the "noisy patient data with known causative genes" dataset. The x-axis is the threshold for the causative gene rank. The y-axis is the ratio of patients satisfying the ranking threshold.

On the "noisy patient data with known causative genes" dataset, we tested all the five method on a set of $2,488$ available genes that have at least one HPO term annotation. The result shows that $PhenoSim$ performed the best in all the five methods (Figure 2). On 86.72% simulated patients, their causative genes are ranked the highest when $PhenoSim$ is applied. In comparison, the percentages of the highest ranked causative genes using Masino, Lin, Jiang, and Schlicker methods are 77.69%, 37.92%, 17.67% and 49.26% respectively. On 98.48% of simulated patients, the causative genes are ranked among top 10 using $PhenoSim$, while the percentages using Masino, Lin, Jiang, and Schlicker methods are 97.12%, 87.69%, 34.89% and 91.97% respectively. Furthermore, Figure 2 shows that the causative gene constantly ranks significant higher on $PhenoSim$ than on the other methods if a high-rank threshold (r) is applied. It indicates that $PhenoSim$ could be potentially helpful to narrow down the causative gene candidate set in practical clinical studies.

C. Performance evaluation on disease prediction

For a given patient, we computed the similarity score between every disease and the patient using $PhenoSim$, and then rank all the diseases by their similarity scores from the largest to the smallest. If the patient-associated disease's rank is higher than any other genesis, we conclude that $PhenoSim$ can accurately predict the disease of the patient.

On the "noisy patient data with known disease" dataset, we tested all the five methods on $2,552$ diseases appeared in both HPO and OMIM. The result shows that $PhenoSim$

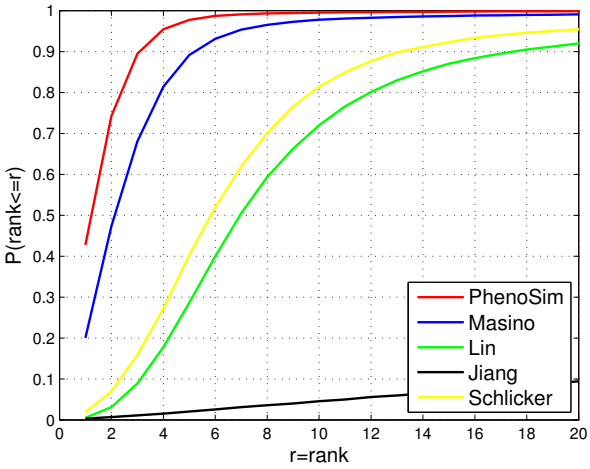

Fig. 3: Cumulative distribution of the rank of the patient-associated diseases on the "noisy patient data with known diseases" dataset. The x-axis is the threshold for the disease rank. The y-axis is the ratio of patients satisfying the ranking threshold.

performed the best in all the five methods (Figure 3). The patient-associated diseases are ranked the highest on 42.74% of the patients if $PhenoSim$ is applied. In comparison, the percentages using Masino, Lin, Jiang, and Schlicker methods are 20.04%, 0.50%, 0.32% and 1.82% respectively.

IV. Conclusion

Recently, next generation sequencing techniques have significantly accelerated disease diagnosis. However, for many diseases with complex phenotypes and high genetic heterogeneity, the disease diagnosis remains challenging. Hence, HPO-based phenotype similarity could be a powerful tool to effectively accelerate the disease diagnosis process. In this article, we proposed a novel method called $PhenoSim$ to measure the phenotype semantic similarity by using a path-constrained Information Content-based method. By well-modeling the noises in patient phenotype datasets, $PhenoSim$ outperforms four existing approaches on all the four patient datasets on causative gene prediction and disease prediction.

Acknowledgment

This project was supported by the National Natural Science Foundation of China (Grant No. 61332014, 61272121); Chemical Sciences, Geosciences and Biosciences Division, Office of Basic Energy Sciences, Office of Science, U.S. Department of Energy (Grant No. DEFG02-91ER20021); U.S. National Science Foundation (Grant No. 1458556); the Northwestern Polytechnical University (Grant No. G2016KY0301); the Fundamental Research Funds for the Central Universities (Grant No. 3102016QD003); and the National High Technology Research and Development Program of China (Grant No. 2015AA020101, 2015AA020108, 2014AA021505).

References

[1] 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.

[2] A. R. Deans, S. E. Lewis, E. Huala, S. S. Anzaldo, M. Ashburner, J. P. Balhoff, D. C. Blackburn, J. A. Blake, J. G. Burleigh, B. Chanet *et al.*, "Finding our way through phenotypes," *PLoS Biol*, vol. 13, no. 1, p. e1002033, 2015.

[3] J. A. Cruz, L. J. Savage, R. Zegarac, C. C. Hall, M. Satoh-Cruz, G. A. Davis, W. K. Kovac, J. Chen, and D. M. Kramer, "Dynamic environmental photosynthetic imaging reveals emergent phenotypes," *Cell Systems*, vol. 2, no. 6, pp. 365–377, 2016.

[4] I. Kahanda, C. Funk, K. Verspoor, and A. Ben-Hur, "Phenostruct: Prediction of human phenotype ontology terms using heterogeneous data sources," *F1000Research*, vol. 4, 2015.

[5] D. Smedley, J. O. Jacobsen, M. Jäger, S. Köhler, M. Holtgrewe, M. Schubach, E. Siragusa, T. Zemojtel, O. J. Buske, N. L. Washington *et al.*, "Next-generation diagnostics and disease-gene discovery with the exomiser," *Nature protocols*, vol. 10, no. 12, pp. 2004–2015, 2015.

[6] S. Köhler, M. H. Schulz, P. Krawitz, S. Bauer, S. Dölken, C. E. Ott, C. Mundlos, D. Horn, S. Mundlos, and P. N. Robinson, "Clinical diagnostics in human genetics with semantic similarity searches in ontologies," *The American Journal of Human Genetics*, vol. 85, no. 4, pp. 457–464, 2009.

[7] 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.

[8] Y. Deng, L. Gao, B. Wang, and X. Guo, "Hposim: an r package for phenotypic similarity measure and enrichment analysis based on the human phenotype ontology," *PloS one*, vol. 10, no. 2, p. e0115692, 2015.

[9] N. L. Washington, M. A. Haendel, C. J. Mungall, M. Ashburner, M. Westerfield, and S. E. Lewis, "Linking human diseases to animal models using ontology-based phenotype annotation," *PLoS Biol*, vol. 7, no. 11, p. e1000247, 2009.

[10] 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.

[11] 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.

[12] Z. Teng, M. Guo, X. Liu, Q. Dai, C. Wang, and P. Xuan, "Measuring gene functional similarity based on group-wise comparison of go terms," *Bioinformatics*, p. btt160, 2013.

[13] R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos, "Phenomenet: a whole-phenome approach to disease gene discovery," *Nucleic acids research*, vol. 39, no. 18, pp. e119–e119, 2011.

[14] C. Pesquita, D. Faria, H. Bastos, A. Falcão, and F. Couto, "Evaluating go-based semantic similarity measures," in *Proc. 10th Annual Bio-Ontologies Meeting*, vol. 37, no. 40, 2007, p. 38.

[15] X. Zhou, J. Menche, A.-L. Barabási, and A. Sharma, "Human symptoms–disease network," *Nature communications*, vol. 5, 2014.

[16] T. Opsahl, F. Agneessens, and J. Skvoretz, "Node centrality in weighted networks: Generalizing degree and shortest paths," *Social networks*, vol. 32, no. 3, pp. 245–251, 2010.

[17] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online mendelian inheritance in man (omim), a knowledge-base of human genes and genetic disorders," *Nucleic acids research*, vol. 33, no. suppl 1, pp. D514–D517, 2005.

[18] J. OMadadhain, D. Fisher, P. Smyth, S. White, and Y.-B. Boey, "Analysis and visualization of network data using jung," *Journal of Statistical Software*, vol. 10, no. 2, pp. 1–35, 2005.

[19] D. Lin, "An information-theoretic definition of similarity." in *ICML*, vol. 98. Citeseer, 1998, pp. 296–304.

[20] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *arXiv preprint cmp-lg/9709008*, 1997.

[21] A. Schlicker, F. S. Domingues, J. Rahnenführer, and T. Lengauer, "A new measure for functional similarity of gene products based on gene ontology," *BMC bioinformatics*, vol. 7, no. 1, p. 1, 2006.