Challenges of Trustable AI and Added-Value on Health
B. Séroussi et al. (Eds.)

doi:10.3233/SHTI220611

Integration of Annotated Phenotype, Gene and Chemical Text Data to Advance Exposome Informatics

Christopher HAWTHORNE[a] and Guillermo H. LOPEZ-CAMPOS[a,1]
[a] *Wellcome-Wolfson Institute for Experimental Medicine, Queen's University Belfast, Belfast, UK*

Abstract. The field of phenomics has a range of biomedical informatics tools such as the Human Phenotype Ontology, providing a structured vocabulary with relationships between abnormal phenotype terms. Artificial intelligence has been widely used for entity extraction and tagging large corpora of text from PubMed and is reflected in applications such as PheneBank and PubTator. Phexpo is a tool for predicting chemical – phenotype relationships and vice-versa, although lacks the ability to decipher known relationships from unknown. Integration of these three resources can provide new meaningful relationships between phenotypes, genes and chemicals and has yet to be fully leveraged. Here we present a methodology to construct two new datasets for phenotype – gene and phenotype – chemical relationships and showcase how these datasets can be used to enhance exposome informatics.

Keywords. Phenotypes, Chemicals, Text-mining, Data Integration, Bioinformatics.

1. Introduction

Diseases and their relationship and interactions with human health is a continual research question. Diseases can routinely be broken down into their constituent phenotypes (the observable physical characteristics) in the example of COVID-19, which at the start of the pandemic was characterized by a continuous cough, fever and fatigue. Phenotypes and their individual relationships to other entities such as chemicals or genes is of great biomedical informatics interest. The Human Phenotype Ontology (HPO) provides a gold standard ontology of phenotype terms as well as associated genes derived from disease-phenotype associations. Text-mining resources such as PheneBank [1] sought to annotate the entirety of PubMed with phenotypic annotations from HPO, other resources which annotate the entirety of PubMed are PubTator [2] which provides genes and chemical annotation. Phexpo [3], is a previously published tool by the group which predicts potential chemical and phenotype relationships using their overlapping genes. Although this tool provides chemical and phenotype predictions, the results have no clear indication if relationships are known or unknown in the scientific literature and have to be validated through manual literature retrieval. The leveraging and integration of these resources has yet to be fully realized. Hence, we utilize these resources to present a new methodology to construct phenotype-gene relationships and phenotype-chemical relationships utilizing gold standard text-mining data from PheneBank and PubTator to further expand phenotype-gene and phenotype-chemical relationships to compare to

HPO's phenotype gene dataset and integrate them with phexpo to validate predictions as well as highlight novel relationships respectfully to further exposome informatics.

2. Methods

PheneBank data was preprocessed to filter out annotation scores greater than the 1st quartile based on first annotations and joined with PubTator gene or PubTator chemical datasets (2020-02-15) in R based on PubMed ID. A further quality control step was applied considering only phenotype – gene relationships sharing at least 3 PubMed IDs. An example phenotype of 'Osteoporosis' and a FDR p.value adjusted threshold of 0.05 were used for phexpo and the phenotype – chemical dataset was compared to the results.

3. Results

Phenotype – gene joining produced 12,272 phenotypes associated to genes. Compared against HPO's phenotype-gene dataset (2021-02-08), we found 1872 phenotypes that were not present and could be associated to genes based on co-occurrence (shared the same PMID). Additionally, we extracted 5,928,876 pairs of phenotype – chemical literature co-occurrences and when these were compared with the results from phexpo where we had an overlap of 611 chemicals, 1,183 chemicals significantly predicted in phexpo only and 2,092 relationships only occurring in the literature.

4. Conclusions

We have shown a new methodology for a novel data integration approach of phenotype-gene and phenotype-chemical interactions, this has facilitated the expansion of the relationships among these concepts. The phenotype-chemical relationships have been combined with the results generated using phexpo validating some of the relationships and highlighting those relationships identified by phexpo that may have not been explored yet in the literature, as phexpo has the potential to predict chemical associations for over 5,000 phenotypes, this provided an approach to validate known relationships and highlight unstudied relationships on a major scale.

References

[1] Pilehvar MT, Bernard A, Smedley D, Collier N. PheneBank: a literature-based database of phenotypes. Bioinformatics. 2021. doi:10.1093/bioinformatics/btab740

[2] Wei CH, Kao HY, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Res. 2013;41: 518–522. doi:10.1093/nar/gkt441

[3] Hawthorne C, Simpson DA, Devereux B, López-Campos G. Phexpo: a package for bidirectional enrichment analysis of phenotypes and chemicals. JAMIA Open. 2020;3: 173–177. doi:10.1093/jamiaopen/ooaa023