Phenotype Information Retrieval for Existing GWAS Studies

Neda Alipanah, Ph.D., Ko-Wei Lin, Ph.D., Vinay Venkatesh, Seena Farzaneh, Hyeon-eui Kim, RN, Ph.D.
Division of Biomedical Informatics, School of Medicine
University of California San Diego La Jolla, CA, USA
nalipanah@ucsd.edu

Abstract

The database of Genotypes and Phenotypes (dbGaP) is archiving the results of different Genome Wide Association Studies (GWAS). dbGaP has a multitude of phenotype variables, but they are not harmonized across studies. We propose a method to standardize phenotype variables by classifying similar variables based on semantic distances. We first extracted variables description, enriched them using domain knowledge, and computed the distances among them. We used clustering techniques to classify the most similar variables. We used domain experts to audit clusters, annotated the clusters with appropriate labels, and used re-clustering to build a semantically-driven Genotypes and Phenotypes (sdGaP) ontology using the UMLS semantic network and metathesaurus. The sdGaP ontology allowed us to expand user queries and retrieve information using a semantic metric called density measure (DM). We illustrated the potential improvement of information retrieval using the sdGaP ontology in one search scenario using the variables from the Cleveland Family Study.

Introduction

Phenotypes are observable physical or biochemical characteristics of an organism, as determined by both genetic variants and environmental influences. Genome-Wide Association Studies (GWAS) investigate many common genetic variants to check whether they are associated with diseases. Several researchers have produced interesting findings with the help of phenotype mapping and harmonization [7,8,9], including PheWAS [4]. In PheWAS, the association between a number of common genetic variations and a wide variety of phenotypes are systematically characterized. The database of Genotypes and Phenotypes (dbGaP) is the main repository of genotypes and phenotypes. The problem with phenotype variables in dbGaP is that they are not harmonized and semantically equivalent variables have several identifiers that make it almost impossible to effectively find them accurately. Furthermore, semantic relations between phenotype variables are not specified in dbGaP. Thus, it is not possible to retrieve variables semantically based on researcher queries. To achieve this goal, we propose to structure phenotypes in dbGaP using custom built software that retrofits existing data sets into information models. This information model based representation formalizes the semantics of phenotype variables. We propose to develop an information model semi-automatically, based on expert human review and NLP-based cluster analysis of variable attributes. Later, we use the semantic hierarchy for expanding and retrieving answers to researcher queries. We use a semantic metric called Density Measure (DM) for information retrieval. The rest of this paper is organized as follows. In section 1, we present our method to solve the stated problem including the technical details of classifying phenotype variables. In section 2, details of sdGaP ontology development are presented. In section 3, we explain how the sdGaP
ontology affects information retrieval. In section 4, we present preliminary experimental results in a selected query within the variables from the Cleveland Family Study (dbGaP Study Accession: phs000284.v1.p1). Finally, the conclusion is presented in section 5.

1. Methods
To build a phenotype data model for dbGaP, namely sdGaP, we computed the semantic hierarchy of dbGaP variables using data mining techniques and background domain knowledge, as depicted in Figure 1. That is, we used the UMLS semantic network as the basis for our sdGaP ontology. We classified dbGaP variables into clusters and domain experts cleaned up these clusters. They relocated the variables that were incorrectly grouped into certain clusters. They then combined some existing clusters or created new clusters. Later, we assigned labeled clusters to the corresponding concepts in the UMLS Metathesaurus and updated sdGaP ontology. Section 1.1 below explains the details of operations for clustering using distances. We computed the semantic hierarchy between phenotype variables as described in section 1.2.

1.1. String-based Variables Distance Calculation
To calculate distances between variables in dbGaP, first we needed to extract the properties of phenotype variables. Variables in dbGaP have different properties including Name, Description, Type, Unit, and (Max-Min) values. We only used variable descriptions to compute the distance between phenotype variables. For this, we built a distance matrix to save the distances. First we expanded variable descriptions using the UMLS-MetaMap. MetaMap is a highly configurable application that maps biomedical text to the UMLS Metathesaurus. MetaMap returned the relevant CUI (Concept Unique Identifiers) and associated semantic types as expanded descriptions of phenotype variables. Then, we calculated the distances between the expanded descriptions using a Vector Space Model [6]. The vector space model is an algebraic model for representing variable description as vectors of identifiers. Later, we calculated the cosine similarity between variable vectors and computed the distances between them. Similar variables had distance equal to 0. We built Distance Matrix and used it to cluster variables.

1.2. Semantic Hierarchy Extraction on dbGaP Variables
As discussed earlier, we employed domain experts (HK and KL) to clean up the clusters. Then, we assigned the clusters from section 1.1 to the concepts of the UMLS Metathesaurus. We extracted semantic hierarchy details for each cluster using UMLS Metathesaurus sub networks. First we built the distance matrix and computed all variables related to each UMLS concepts (e.g. “Diseases”). Then, we considered each cluster to find possible subclasses for the next levels of hierarchy. If the number of elements was sufficient (i.e. greater than 3) at each classification level for a new cluster, we recursively produced the next level clusters, and mapped them to a network of UMLS Metathesaurus. Different groups of clusters were considered as instances in the sdGaP ontology. Figure 1 presents a small example of the taxonomy. Variable “phv00123020.v1” comes from the variable description “CVD: self report of MD dx of cvd”, and “phv00123021.v1” comes from the variable description “CVD: self report of MD dx of cvd (missing recoded as no)”. These variables are semantically located in cardiovascular cluster.
2. Classification and Ontology Creation
To develop the sdGaP ontology, we used the UMLS semantic network as the base ontology and appended UMLS metathesaurus terms as subclasses of the sdGaP base ontology. Later, we assigned phenotype variables in each cluster to the corresponding concepts in sdGaP as instances. The instantiation was done only for low-mid level classes in the sdGaP ontology and not the high level because we needed to index phenotype variables to the exact specific classes of the sdGaP ontology. For example, we assigned the “Disease or Condition” cluster to the “Disease” class in the sdGaP ontology. Further, we created the “Cardiovascular Disease” subclass. At the next level, “Heart Disease” was created. Finally, we instantiated variables with id=”phv00123020.v1” and “phv00123021.v1” to “Cardiovascular Disease” class in sdGaP and other relevant variables as shown in Figure 1 to the “Heart Disease” class.

3. sdGaP Information Retrieval
The sdGaP ontology was built to enable researchers to get accurate, semantically relevant variables based on their query keywords. For this, we expanded user query keywords in the sdGaP ontology using a semantic metric called Density Measure (DM).

Density Measure (DM): This measurement expands some details of each query keyword in the sdGaP ontology graph using parent-children relations. Suppose \( q_j \) is the query keyword and \( e \in sdGaP \) is a class in sdGaP and \( e = StringMatch(q_j) \). Query expansion includes all entities that have children relation with \( e \), that is, \( \{e_1, e_2, ..., e_m\} \)[3].

4. Preliminary Results
In this section, we present the results of our proposed query expansion method on the sdGaP ontology. We used 5 data sets and 2,339 phenotype variables as our dataset from the Cleveland Family Study (CFS). First we expanded variable descriptions using UMLS-MetaMap and calculated the distances between variables using the method described in section 1.1. Second, we performed cluster analysis using the open source machine learning tool Weka with the X-mean clustering algorithm [2]. We used domain experts to clean up cluster elements for the next level of the hierarchy to build the
baseline semantic network in sdGaP. Then we mapped labeled clusters to the sdGaP classes. The X-mean clustering resulted in 35 clusters for relevant variables. Our domain experts merged some clusters, created new clusters and produced 23 top level variable clusters. Figure 2 presents the distribution of phenotype variables in clusters after domain expert review and categorization at the first level of semantic hierarchy. We repeated the clustering by expanding variable descriptions using the UMLS subnetworks and eventually built a pilot sdGaP ontology. For the next levels, as presented in the Figure 1, we used UMLS subnetworks (e.g., disease subnetworks including “Heart Disease”, …,”Cardiovascular” diseases) and generated other levels of the sdGaP ontology hierarchy by clustering disease variables within the Cleveland Family Study (phs000284.v1.p1).

| Cardiovascular Disease | Heart Disease |
|------------------------|--------------|
| phv00123495.v1, phv00122277.v1, phv00122290.v1, phv00122286.v1, phv00122274.v1, phv00122294.v1, phv00122281.v1, phv00123500.v1, phv00123020.v1, phv00123021.v1, CVD: self report of MD dx of cvd (missing recoded as no) phv00123017.v1, phv00122284.v1, phv00122285.v1, phv00123484.v1, phv00122293.v1, phv00122289.v1, phv00122280.v1, phv00122292.v1, phv00122283.v1, phv00122273.v1, phv00123018.v1, phv00123016.v1, phv00123498.v1, phv00122288.v1, phv00122269.v1, phv00122272.v1, phv00123022.v1, phv00123019.v1, phv00122246.v1, phv00124274.v1, phv00124291.v1, phv00124287.v1, phv00124283.v1, phv00124313.v1, phv00124261.v1, phv00124310.v1, phv00124280.v1, phv00124275.v1, phv00124262.v1, phv00124293.v1, phv00124268.v1, phv00124279.v1, phv00124302.v1, phv00124278.v1, phv00124304.v1 | phv00122390.v1, Hospitalized for heart disease phv00122277.v1, Irregular heart beat diagnosed by doctor (A) phv00122286.v1, Other heart disease diagnosed by doctor (AC) phv00122274.v1, Heart attack diagnosed by doctor (A) phv00122281.v1, Heart failure diagnosed by doctor phv00122270.v1, Angina diagnosed by doctor phv00123020.v1, CVD: self report of MD dx of cvd (missing recoded as no) phv00122284.v1, Other heart disease (AC) phv00122285.v1, Specify heart disease phv00122280.v1, Heart failure (A) phv00122283.v1, Heart failure problem still present phv00122273.v1, Heart attack (A) phv00122288.v1, Other heart disease still present (AC) phv00122269.v1, Angina (A) phv00122272.v1, Angina still present (A) phv001222460.v1, Specify heart condition phv00124291.v1, DFVD* - Diagnosed Other Heart Dis. phv00124261.v1, DFVD* - Diagnosed Chest Pain from a heart condition diagnosed phv00124262.v1, DFVD* - Diagnosed Coronary Angioplasty phv00124293.v1, DFVD* - Diagnosed Congestive Heart Failure phv00124268.v1, DFVD* - Diagnosed Coronary Bypass diagnosed phv00124279.v1, DFVD* - Diagnosed Irregular Heart Beat phv00124302.v1, DFVD* - Diagnosed Implant of Cardiac Pacemaker diagnosed phv00124278.v1, DFVD* – Diagnosed Peripheral Artery Disease of Claudication of the Legs |

Table 1: Gold Standard of Cardiovascular and Heart Disease in phs000284.v1.p1 Study. Highlighted items were included by automated clustering. *DFVD=Date From Visit Day

Table 1 presents the gold standard for cardiovascular and heart disease in the phs000284.v1.p1 study. The highlighted variables are the ones that appear in our second and third level clusters without domain expert review. Using the sdGaP ontology for the “cardiovascular disease” query, we know that heart disease is a child of cardiovascular disease. If we search for “cardiovascular disease” in the ontology, we retrieve only two variables “phv00123020.v1”, “phv00123021.v1” which makes our recall to be 2/45=0.04. When we expand query using the DM semantic measure, we retrieve all highlighted variables of cardiovascular and heart disease, which results in recall= 18/45= 0.4.

5. Conclusion
We outlined an ontology based information retrieval approach based on our sdGaP ontology. This is achieved by a semi-automatic ontology creation process. Domain expert’s manual review and refinement of the clusters generated through algorithmic clustering method was an important factor for the accuracy of our work. Through this exploratory work, however, we discovered that our method is not scalable for very large data, both for distance calculation between variables and UMLS hierarchy extraction. Parallel processing approaches will be investigated in future work.

**Acknowledgement**

This work was supported by the grants UH2HL108785 (NHLBI) and R01HS019913 (AHRQ) under supervision of Dr. Lucila Ohno-Machado.

![Distribution of Phenotype Variables in Different Categories](image)

**Figure 2:** Categorization of Cleveland Family Study Variables

**References**

1. Aronson A. [2012], The UMLS Metathesaurus, MetaMap Portal [Online], Available: http://metamap.nlm.nih.gov/.
2. Frank E., Holmes G., Mayo M., Pfahringer B., Smith T. Witten I. [2012], Weka 3: Data Mining Software in Java, Available: http://www.cs.waikato.ac.nz/ml/weka/.
3. Alipanah N., Khan L., Thurasingham B., Optimized ontology-driven query expansion using map-reduce framework to facilitate federated queries, International Journal of Computer Systems Science and Engineering, Volume 27 Number 2 March 2012.
4. Pendergrass A., Brown-Gentry K., Dudek S. M., Torstenson E. S., Ambite J. L., Avery C. L., Buyske S., Cai C., Fesmeyer M. D., Haiman C., Heiss G., Hindorff L. A., Hsu C. N., Jackson R. D., Kooperberg C., Le Marchand L., Lin Y., Matise T. C., More-land L., Monroe K., Reiner A. P., Wallace R., Wilkens L. R., Crawford D.C., and Ritchie M.D. 2011. The use of phenome-wide association studies (phewas) for exploration of novel genotype-phenotype relation ships and pleiotropy discovery. *Genetic Epidemiology*, 35(5):410–422.
5. Plone Foundation [2012], The UMLS Semantic Network in OWL Ontology [Online], Available: http://krono.act.uji.es/people/Ernesto/UMLS_SN/owl.
6. Salton G., Wong A., Yang C.S. A vector space model for automatic indexing, Communications of ACM, Volume 18 Issue 11, Nov.1975:613-620.
7. Johnson AD, O’Donnell CJ. An open access database of genome-wide association results, BMC Med Genet.2009:10:6.
8. Ikram MA, Seshadri S, Bis JC, et al. Genomewide association studies of stroke. *N Engl J Med*. Apr 23 2009;360(17):1718-1728.
9. Zeggini E, Ioannidis JP. Meta-analysis in genome-wide association studies. Pharmacogenomics. Feb 2009;10(2):191-201.