UMLS-KG: A biomedical knowledge graph built into the UMLS

Ben Stear[1], MS; Taha Mohseni Ahooyi[1], PhD; Shubha Vasisht[1]; Jonathan Silverstein[3,4], MD, MS, FACS, FACMI; Tiffany Callahan[5], PhD; Deanne Taylor[1,2], PhD.

[1]Department of Biomedical and Health Informatics, Children's Hospital of Philadelphia; [2]Perelman School of Medicine, University of Pennsylvania; [3]Health Sciences and Institute for Precision Medicine, University of Pittsburgh; [4]Department of Biomedical Informatics, University of Pittsburgh; [5]Anschutz Medical Campus, University of Colorado Denver

Introduction

Complex genetic diseases could be a consequence of multiple interacting variations that affect gene function. Current annotation pipelines to determine the effect of genetic variation typically measure the deleterious nature of one variation at a time, reflecting Mendelian models of gene inheritance and disease effect. A common approach to studying multi-gene contributions to disease is to use "gene sets" from various sources. Integrated datasets with deeply typed ontological categorizations can provide a richer background for determining multi-genic functional relationships. Using this approach allows for determination of new gene sets related by functional, semantic and categorical links not easily accessible without massive data integration. As a result,we can explore the effects of new multi-gene interactions in complex genetic diseases.

The Unified Medical Language System (UMLS) is a large repository of biomedical ontologies, vocabularies, and relationships. We "bring the data to the ontologies" by mapping quantitative data from various sources into the UMLS by modeling it as a property graph using the Neo4j graph database platform. To date, we've integrated eight additional datasets into the ontology systems native to the UMLS knowledge graph (UMLS-KG) framework including several quantitative datasets and gene-to-phenotype mappings across and between mouse and human genomes. This knowledge graph implementation allows for fast, complex queries across a wide range of biomedical terms and quantitative data. Our integration has produced a knowledge graph with approximately 28 million nodes and 76 million relationships. We discuss the resulting graph characteristics and the query results from this massively complex ontological, categorical and multi-omics data integration.

Background

The Unified Medical Language System (UMLS) is a large data repository created by the National Library of Medicine (NLM) in 2004 which consists of over 60 controlled biomedical vocabularies and ontologies. The UMLS is updated multiple times a year and was created "to overcome two significant barriers to effective retrieval of machine-readable information' : the variety of names used to express the same concept and the absence of a standard format for distributing terminologies" [1]. The schema for the UMLS works by grouping synonymous terms under umbrella 'Concepts'. These UMLS Concepts allow for multiple biomedical researchers to refer to a common UMLS Concept even if they are all using different (but synonymous) terms from separate ontologies or vocabularies. In addition to the relationships within individual ontologies/vocabularies, the UMLS contains relationships between ontologies/vocabularies created by the UMLS curators. Recently, a property graph implementation of the UMLS (UMLS-KG) has been developed using the Neo4j platform by researchers at the University of Pittsburgh.

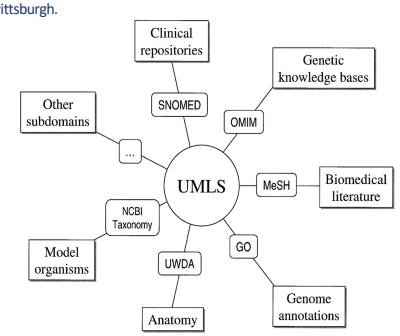


Figure 1. Examples of the ontologies and vocabularies that exist within the Unified Medical Language System (UMLS).

Methods

We integrated eight additional datasets into the UMLS-KG. We began by adding the **Mammalian Phenotype** (MP) which is an ontology that describes the hierarchy of mouse phenotypes. The MP data was directly loaded into a Neo4j graph instance using the Neosemantics tool from Neo4j. This graph was then extracted, formatted to fit the UMLS-KG schema and loaded into the UMLS-KG. The MP consists of ~13,200 phenotypes and their relationships. We then integrated gene expression per tissue data from the **Genotype-Tissue Expression (GTEx)** Portal. From GTEx, we added the gene expression per tissue dataset and the expression quantitative trait loci (eQTL) dataset into the UMLS-KG. The gene expression dataset contains expression levels from 54 tissues and 56,200 genes (transcripts). The eQTLs dataset contains over 1.2 million eQTLs from 49 tissues. We created Concept nodes for each eQTL and each tissue-gene expression pair. The eQTL Concepts were then connected to their corresponding tissue node (UBERON), gene node (HGNC) and variant node (dbSNP). The gene expression nodes are connected to their corresponding tissue node and gene node. We also integrated quantitative data from GTEx including P-values for the eQTL data and transcripts per million (TPM) for the gene expression data. **Gene annotation** data was obtained from the GENCODE website. From this dataset we incorporated gene location (start and end positions), chromosome and strand (+ or -). This data is connected to the respective HGNC Concept. **Single cell RNA-seq** data was obtained from the Asp et al. 2019 paper [2] and was analyzed and clustered using Seurat. These nodes were connected to their respective Cell Ontology and HGNC nodes. We also integrated gene-to-pathway data from MSigDB, variants associated with disease from ClinVar, and chemical-to-gene expression from LINCs.

Name of Dataset	Type of Data	Dat points
Mammalian Phenotype (MP)	Ontology	13,240
Gene-Tissue Expression Portal (GTEx)	Quantitative (Gene-Tissue-Expression, eQTLs)	2,993,782
Developmental Heart Single Cell RNAseq (Asp et al.)	Quantitative (Gene-Celltype-Expression)	2,816
HGNC Gene Annotations (GENECODE)	Annotations	75,044
Human-Mouse Ortholog mappings (HCOP) – Mouse Gene IDS	Relationships	27,070
Mouse Genotype-Phenotype Mappings (IMPC, MGI)	Relationships	505,994
Human Genotype-Phenotype Mappings (MONARCH)	Relationships	1,762,468
Human-Mouse Phenotype mappings (PheKnowLator)	Relationships	2,504
LINCS (L1000, Cmap + KINOMEscan + KiNativ)	Relationships	5,036,182
MsigDB (pathway collections)	Relationships	1,391,721
ClinVar	Relationships	21,890
	Total	**11,832,711**

Table 1. Ontologies, datasets and mappings integrated into the UMLS-KG.

Human-to-Mouse orthologs were obtained from the HGNC Comparisons of Orthology Predictions (HCOP) tool. The UMLS does not contain any mouse gene data so we created new mouse gene Concepts and mapped them using the HCOP data to their corresponding human ortholog. Out of the 41,638 HGNC codes in the UMLS, the HCOP tool found at least one mouse ortholog for 20,715 HGNC codes. **Mouse genotype-to-phenotype mappings** were obtained from multiple datasets from two databases, namely, the international mouse phenotyping consortium (IMPC) and the mouse genome informatics (MGI) database. The datasets from IMPC and MGI were combined to create a master genotype-to-phenotype dataset. This master dataset contains 10,380 MP terms that are mapped to at least one gene and 17,936 genes that are mapped to at least one MP term. **HPO-MP mapping** data that connects human phenotypes to mouse phenotypes was generated using the PheKnowLator tool [2]. We chose to map only phenotypes related to congenital heart defects as that is our initial use case. The mappings that PheKnowLator generated were then checked and edited manually for accuracy. We kept only the highest quality mappings which left us with ~1,000 mappings.

Results

We integrated eight datasets into the UMLS-KG, a knowledge graph model of the UMLS implemented in Neo4j. We've added over 10 million new nodes and 30% new relationships to the graph, representing about a 50% increase in nodes and a 36% increase in relationships compared to the original UMLS-KG. The datasets that have been added are associated with genotypic, phenotypic and genetic data (or mappings). We now have these datasets integrated into a powerful knowledge graph containing millions of ontological and semantic mappings which can be leveraged to enrich queries and provide contextual and categorical meaning to biomedical questions.

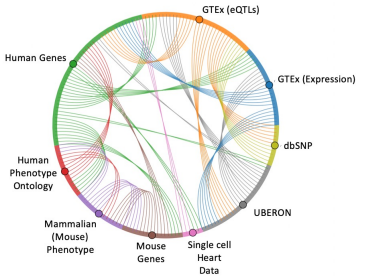


Figure 2. Interconnectedness of data sources we added to the UMLS-KG. *Note the dbSNP dataset contains variants from the GTEx datasets and does not represent the entirety of dbSNP.

For example, a researcher may want to find all genes that are associated with a family of phenotypes (such as congenital heart defects, HP:012345) that have evidence from a mouse gene knockout experiment. Without our knowledge graph you'd need to find the mammalian phenotype equivalents for all human phenotypes you are interested in, then find all mouse genes associated with each mammalian phenotype and then find all human orthologs for these mouse genes. Using our graph, this question can be answered easily and quickly with a single query. Furthermore, a researcher might want to know all the eQTLs in genes associated with a phenotype that are highly expressed in the gut and the brain and located on chromosome X or Y. Again, without our graph this is a multistep process. You'd need to first find all genes associated with the phenotype of interest. Then find all genes that are highly expressed in the gut and brain, and then find all genes located on chromosome X or Y. Next you'd need to find the union of step 1,2,3 and lastly, find all eQTLs for the resulting gene list. However, with our knowledge graph a question like this can be answered with a single query.

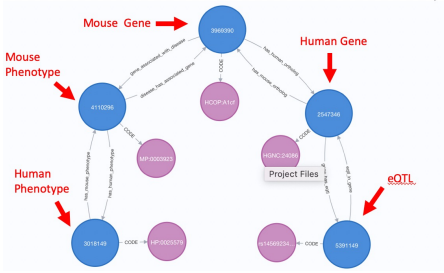
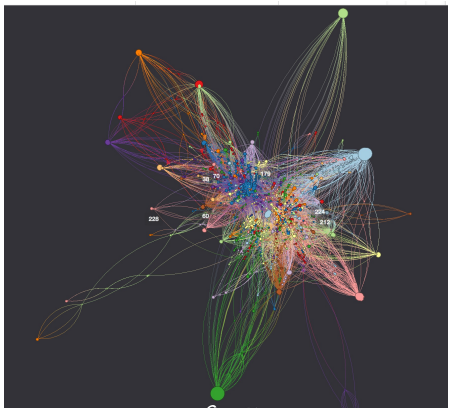


Figure 3. Graphical representation of one of our main use cases. Concept nodes are in blue and their corresponding Code nodes in purple.

Conclusions

Our knowledge graph enables users to query and answer non-trivial biomedical questions quickly and easily. By "bringing the data to the ontologies" we have created a resource

We are currently working on integrating additional datasets such as pathway ontologies,drug-gene interactions, -omics datasets, etc. We are also developing an API around the knowledge graph to parameterize the queries so that researchers do not need to understand the graph schema or the query language to use the UMLS-KG.



References

1. Olivier Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology, *Nucleic Acids Research*, Volume 32, Issue suppl_1, 1 January 2004, Pages D267–D270, https://doi.org/10.1093/nar/gkh061
2. Callahan TJ, Tripodi IJ, Hunter LE, Baumgartner WA. A Framework for Automated Construction of Heterogeneous Large-Scale Biomedical Knowledge Graphs. 2020; BioRxiv DOI: https://doi.org/10.1101/2020.04.30.071407
3. Asp M, Giacomello S, Larsson L, Wu C, Fürth D, Qian X, Wärdell E, Custodio J, Reimegård J, Salmén F, Österholm C, Ståhl PL, Sundström E, Åkesson E, Bergmann O, Bienko M, Månsson-Broberg A, Nilsson M, Sylvén C, Lundeberg J. A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. Cell. 2019 Dec 12;179(7):1647-1660.e19. doi: 10.1016/j.cell.2019.11.025. PMID: 31835037.

Acknowledgements

I want to thank Shubha Vasisht, Taha Moseniah and Deanne Taylor for help and guidance throughout the duration of the project.

This work was supported by an R03 grant from the NIH,

.

Project Github: https://github.com/TaylorResearchLab/CFDIKG