

Utilizing Semantic and Contextual Measures to Evaluate the Similarity between Disease Concepts

Karan Luthria¹, Maricel Kann¹, Olivier Bodenreider²

¹University of Maryland, Baltimore County, Baltimore, MD; ²National Library of Medicine, National Institutes of Health, Bethesda, MD

Abstract:

The critical factor for expensive healthcare is the cost of drug discovery. One possible approach to reduce these expenses is to repurpose drugs based on identifying similarities between diseases through extending disease-variant connections. Current disease network models used to identify potential drug repurposing targets link diseases are based on shared genetic variants. Since genes do not work alone, we have developed an extension to these disease network models that are able to link diseases through protein interactions. To further analyze this disease network, we have investigated hierarchical and contextual measures of similarity between disease concepts. By doing so, we can cluster similar disease phenotypes to obtain more novel links between diseases. Hierarchy based measures involve the use of the well-curated disease hierarchy found in Systematized Nomenclature of Medicine -- Clinical Terms (SNOMED CT). On the other hand, contextual similarity methods involve the use of natural language processing models performed on a text corpus. In this study, we have implemented and compared a series of methods in order to identify areas of significant differences between these two methods which will be utilized to develop a custom disease clustering method that is able to integrate information from SNOMED CT and contextual similarity measures.

Background:

In recent years, the concept of drug repurposing has grown increasingly attractive due to the rising cost (>\$2.5 billion) for new drug development. The process of repurposing an existing drug cost \$300 million on average in a much shorter time period.¹ Current disease network models used to identify potential drug repurposing targets link diseases are based on shared genetic variants. However, genes and gene products (i.e., proteins) commonly interact with other molecules that may contribute to different diseases. Therefore, there is a pressing need to include gene and protein interactions in current models for identifying disease relatedness and potential drug repurposing targets. We have previously combined gene-disease associated data with protein interaction networks to develop an extensive human disease-disease interaction network (HDDN) that can identify similarities between diseases at a complex molecular level. To do so, we utilized a compiled over 80,000 disease-variant associations from various manually curated databases². Since each database contains a unique disease descriptor, we leveraged the Unified Medical Language System (UMLS) to normalize the different terminologies³. However, the UMLS contains a large amount of concepts, many of which are too fine-grained from the identification of unique disease links. There is a critical need to cluster and identify similar disease pairs to extract unique molecular links between disease concepts. For example, we have identified “uninteresting” molecular links between Osteogenesis imperfecta type IV and Osteogenesis Imperfecta. Being able to cluster such similar diseases will aid in the removal of redundant or uninteresting links, thereby making our current disease network more informative. To effectively cluster similar diseases, an accurate disease similarity measure is required. In this study, we have compared the two primary forms of disease similarity, hierarchical based disease similarity, and contextual based disease similarity.

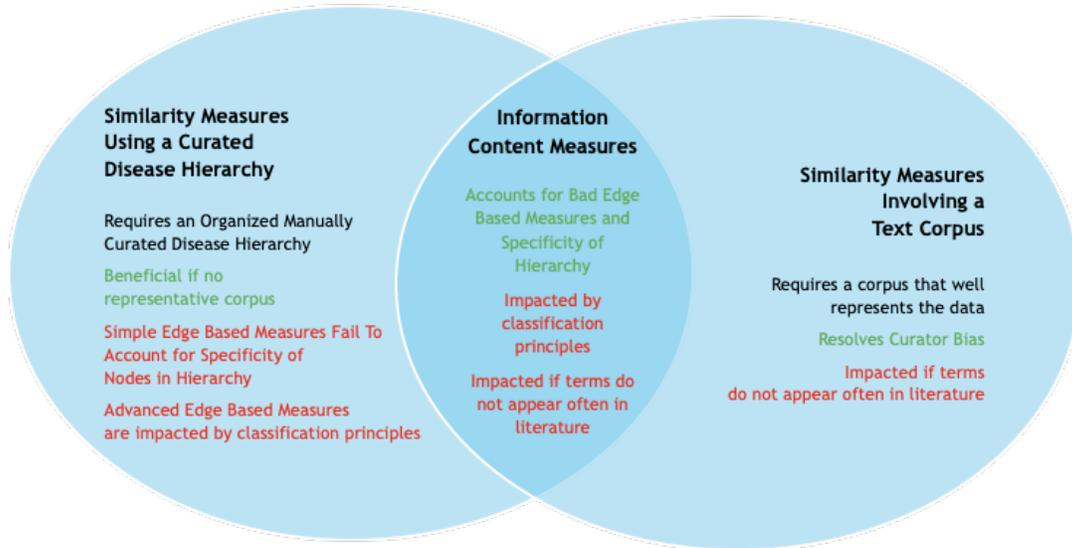


Figure 1: The Venn Diagram depicts the two main umbrellas of determining the similarity between diseases, along with their advantages and disadvantages.

Hierarchy based similarity measures are primarily based on how far two concepts are within the SNOMED CT disease hierarchy. SNOMED CT was chosen to be the ideal disease hierarchy as it is well maintained by SNOMED International, and contains a comprehensive and in-depth collection of disease concepts. Most importantly it is manually reviewed, unlike the computationally complex, unreviewed, disease hierarchy built into the UMLS Metathesaurus. The primary issue with disease hierarchy methods is that they are often subjective to classification principles determined by the curation team. For example, certain disease hierarchies are sorted by body system(s) and thus unable to identify similar diseases between two body systems. Additionally, simple disease hierarchy methods are unable to account for the depth and specificity of concepts within the hierarchy (Fig. 1).

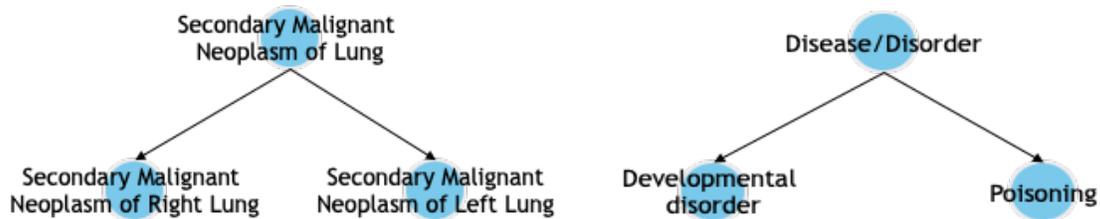


Figure 2: Both Pairs: Developmental Disorder and Poisoning; Secondary Malignant Neoplasm of Right Lung and Secondary Malignant Neoplasm of Left Lung are two edge lengths apart. However, the latter pair is much more similar than the former. This reveals the fundamental flaw behind simple edge based methods.

In order to account for the depth of the hierarchy and the specificity of nodes, Information Content based methods were developed. Such methods involve the use of a corpus. The specificity of nodes is determined by its Information Content (IC), which is defined as the $-\log_2(p)$ ($-\log_2(p)$), the negative log of the probability of encountering a given node in a corpus. The simplest IC based calculation, the Resnik Measure,

defines similarity as the IC of the lowest common ancestor. In Fig. 1, the probability of encountering Secondary Malignant Neoplasm is much lower than the probability of encountering Disease/Disorder, and thus, Secondary Malignant Neoplasm of Left Lung and Secondary Malignant Neoplasm of Right Lung would have a higher Resnik similarity. Recently, advanced edge-based similarity measures were also developed that utilize information such as the number of descents/ancestors a node has to determine its specificity without a corpus. However, such methods assume the hierarchy is evenly spread.

Corpus-based similarity is based on unsupervised machine learning natural language processing models trained on a representative text corpus. Such methods involve determining the similarity of two diseases based on the context around them in the text. Such methods have been previously developed utilizing MEDLINE abstracts and titles as the biological corpus and sentences of UMLS Concepts, or CUIs, as outputted by MetaMap. The primary issue with such methods is that they require a representative text corpus that has a high occurrence of each CUI in literature.

Methods:

1. Preprocessing

To utilize hierarchy based similarity measures, there arises a need to convert from UMLS CUIs to a database with an innate disease hierarchy. From the unique 80,000 disease variant associations we have identified 6723 unique diseases. We noticed that the majority of our disease phenotypes, mapped to UMLS CUIs, originate from OMIM (5951/6723) due to it being a database for inheritable genetic diseases. As a result of OMIM containing no innate disease hierarchy one can utilize, we are required to crosswalk to an alternative database, such as SNOMED CT. The cross-walking to SNOMED CT was performed using a variety of methods. First, we attempted to map to synonymous concepts through the use of various sources. To do so, we primarily utilized the UMLS API to see if there is a direct conversion from UMLS CUI to SNOMED CT Concept. In the event that this was not applicable, we attempted to convert from UMLS CUI to Orphanet Disease Concept, a database for genetic and rare diseases, and then attempted to see if that Orphanet Concept was able to map to SNOMED CT. Secondly, we attempted to map to broader concepts if direct mappings were not applicable. For this, we first attempted to traverse up the UMLS hierarchy until a mapping to SNOMED CT was found, and finally we attempted to lexically truncate the term until it was able to be found within SNOMED CT. Through such methods we were able to convert about 6345/6723 of the CUIs to 2947 Unique SNOMED CT Concepts, the rest of which need to manually mapped.

To utilize contextual based similarity measures, there requires a need for each disease concept to appear in the corpus. Within the 2015 MetaMap baseline, more than $\frac{1}{3}$ of the disease CUIs were not identified in literature. However, upon conversion to SNOMED CT, many of the CUIs converted to broader SNOMED concepts also appeared in the literature. 6178/6723 CUIs are able to convert to SNOMED CT and appear in literature. After doing so, we are then able to proceed to implement a series of measures falling under these two umbrellas for further comparison. More specifically, we would like to identify the main cause between the differences between hierarchy based measures and contextual measures in order to determine an effective similarity measure that is able to take advantage of both sources of information.

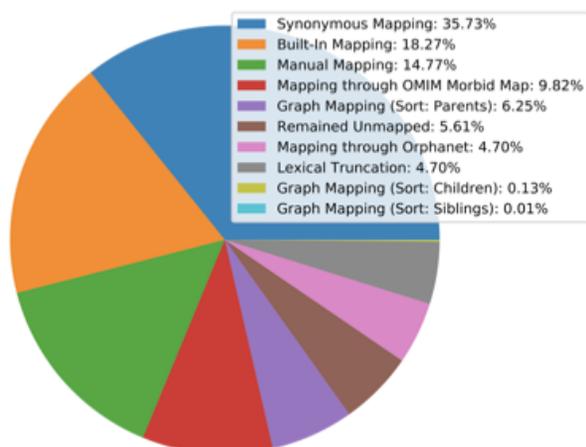


Figure 3: The distribution of methods utilized to map each of the UMLS Disease Concepts to SNOMED CT

2. Implementing Measures

Distance Measure	Utilizing
Edge Distance	Simplistic Hierarchy Only
Pekar and Staab ⁴	Advanced Hierarchy Only
Leacock and Chodorow ⁵	Advanced Hierarchy Only
Betet ⁶	Advanced Hierarchy Only
Resnik ⁸	Information Content
Lin ⁹	Information Content
Jiang and Conrath ⁷	Information Content
SimGIC ¹⁰	Information Content
CUI2Vector Distance ¹¹	MEDLINE Corpus
Co-Occurrence Matrix Cosine Similarity	MEDLINE Corpus

Table 1: The various semantic similarity measures tested and compared within this study.

After preprocessing the data by converting the UMLS disease phenotypes to SNOMED CT concepts, we were now able to implement a series of semantic similarity measures to test and compare. To implement the various hierarchy based measures, we utilized networkx, a python module, for effective information gathering of the disease hierarchy. In order to calculate contextual similarity, we integrated from the already existing word2vec CUI embeddings using CBOW and a Co-Occurrence Matrix from Henry et al. 2017 and Henry et al. 2018 respectively. (CITE BOIS). To calculate information content, we utilized the MetaMap 2018 baseline to identify how often each term appears in MEDLINE Abstracts ranging from 1985-2018. After doing so, we implemented a series of measures as seen in Figure 1. Through doing so, we were

then able to identify areas of significant difference between the corpus based models and the hierarchy based models.

Results and Discussion:

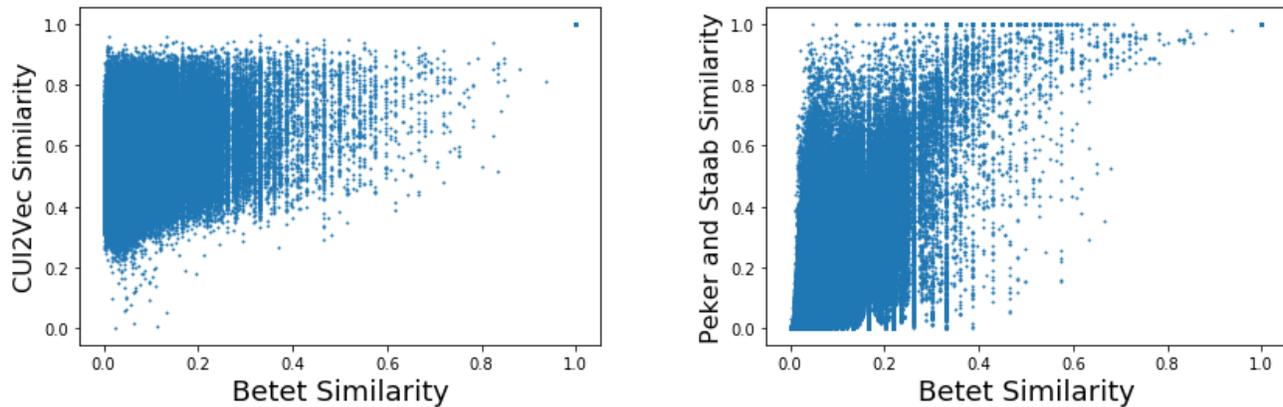


Figure 4: Scatter plots revealing limited correlation between the Betet Similarity (Hierarchy Only) and the CUI2Vec Similarity (Text Corpus Only) (R-squared = 0.056), as compared to two hierarchy only measures (Betet vs Pekar and Staab Similarity) (R-squared = 0.439).

In order to understand and identify if significant differences exist between the hierarchy based and the contextual based measures, we calculated similarity matrices for the 6178 CUIs that were not only identified in literature but also successfully converted into SNOMED CT. This was done utilizing the NIH's High-Performance Computing System known as Biowulf. After computing the various similarity measures, we investigated if there was any significant difference between the CUI2Vec Similarity (a Context Only Based Measure) and the Betet Similarity Measure (a Hierarchy Only Based Measure). To appropriately determine a control, we also compared Betet Similarity to another Hierarchy Only Based Measure, Pekar and Staab. (Cite Bois) As evident in Figure 4, there appears to be no apparent similarity between the two measures from different sources of information. We then determined the sample of data points from which significant disparities were identified between the contextual and the hierarchy based measures. This assisted in identifying missing is-a links between many disease concepts listed under the disease/disorder branch and morphologic abnormalities between these diseases. For example Gastrointestinal stromal sarcoma (morphologic abnormality)(SNOMED: 128756002) and Malignant epithelial neoplasm (disorder) (SCTID: 722688002) have a semantic similarity score of 0, but an extremely high contextual similarity score (0.771651). This is a result of the classification principles used by the SNOMED CT manual curators. There were also significant amount of false positives that occurred as a result of the CUI2Vec model. For example, Glycogen Storage Disease Type IIb and Loeys-Dietz Syndrome are two unrelated disorders, but they had a high CUI2Vec similarity (0.612856). However, the Betet similarity, a hierarchy based method were able to pick up and identify the lack of similarity (0.127060).

Through an analysis of the disease hierarchy semantic similarity measures and the corpus based similarity measures, it is clear that both methods often appear to be more applicable in various scenarios. However, further studying of the SNOMED CT Disease Hierarchy reveals that many of the links that appear to be missing is-a links are a result of the various linking mechanisms in SNOMED CT. For example, the link between Gastrointestinal stromal sarcoma (morphologic abnormality)(SNOMED: 128756002) and Malignant

epithelial neoplasm (disorder) (SCTID: 722688002) is identified to a link identified as an associated morphology link rather than an is-a link.

Conclusions and Future Work:

In order to develop a clean disease similarity measure for clustering, it is clear that neither semantic and contextual similarity measures would be able to satisfy when used alone. The contextual similarity measures were able to identify specific forms of missing is-a links that cannot be found through using the disease hierarchy links. A large portion of these missing links identified by contextual measures can be found through other forms of links in SNOMED CT as they are not an is-a link, but rather a morphology link, a due-to link etc. A need to include such links when identifying disease similarity is critical. As a result, one solution that can be developed is by utilizing disease clusters created through hierarchical clustering of a hierarchy based method, and adding all SNOMED CT concepts that have these additional links to these clusters, as well as the transitive closure of such concepts. Future work to identify methods include these missing links to disease clustering methods is required.

References:

- [1] Y. Cha, et al., "Drug repurposing from the perspective of pharmaceutical companies," *British journal of pharmacology*, vol. 175, no. 2, pp. 168–180, 2018.
- [2] T. A. Peterson, E. Doughty, and M. G. Kann, "Towards precision medicine: advances in computational approaches for the analysis of human variants," *Journal of Molecular Biology*, vol. 425, no. 21, pp. 4047–4063, 2013.
- [3] A. G. Cirincione, K. L. Clark, and M. G. Kann, "Pathway networks generated from human disease phenome," *BMC Medical Genomics*, vol. 11, p. 75, Sep 2018.
- [4] V. Pekar and S. Staab, "Taxonomy learning—factoring the structure of a taxonomy into a semantic classification decision," in *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002.
- [5] C. Leacock, M. Chodorow, C. Fellbaum, et al., "Combining local context and wordnet similarity for word sense identification," *MIT Press*, pp. 305–332, 01 1998.
- [6] M. Batet, D. Sanchez, and A. Valls, "Anontology-based measure to compute semantic similarity in biomedicine," *Journal of biomedical informatics*, vol. 44, no. 1, pp. 118–125, 2011.
- [7] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *arXiv preprint comp-lg/9709008*, 1997.
- [8] P. Resnik, et al., "Using information content to evaluate semantic similarity in a taxonomy," *arXiv preprint comp-lg/9511007*, 1995.
- [9] D. Lin, et al., "An information-theoretic definition of similarity," in *Icml*, vol. 98, pp. 296–304, Citeseer, 1998.
- [10] C. Pesquita, D. Faria, F. M. Couto, et al., "Metrics for go based protein semantic similarity: a systematic evaluation," in *BMC bioinformatics*, vol. 9, p. S4, BioMed Central, 2008.
- [11] S. Henry, C. Cuffy, and B. T. McInnes, "Vector representations of multi-word terms for semantic relatedness," *Journal of biomedical informatics*, vol. 77, pp. 111–119, 2018.