# Towards a More Molecular Taxonomy of Disease

Jisoo Park, Benjamin J. Hescott, and Donna K. Slonim | Department of Computer Science, Tufts University, Medford, MA 02155

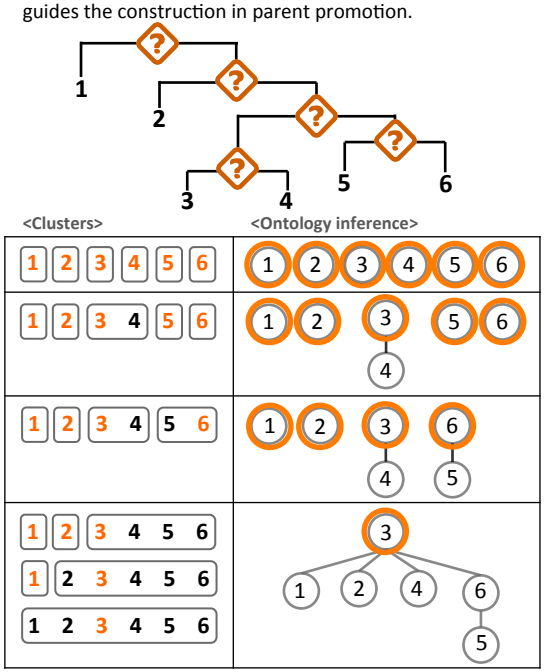
For further information:
jisoo.park@tufts.edu

## Overview

Disease taxonomies have been created for many applications, but they tend not to fully incorporate our growing molecular-level knowledge of disease processes, inhibiting research efforts[1]. In this pilot study, we report the results of several preliminary attempts to infer a hierarchical representation of diseases from disease gene data. Our goal is not to build a new taxonomy using only these data, but to begin to understand the molecular contributions represented in existing taxonomies and how molecular information may be incorporated with existing criteria. Our preliminary results suggest that disease-gene association has the potential to serve as a firm foundation of future representations of the disease landscape.

## Parent Promotion

We cannot use hierarchical clustering alone because internal nodes correspond not to individual diseases but to unnamed sets of diseases. Instead, hierarchical clustering guides the construction in parent promotion.




## Methods

**Disease-gene association** from OMIM + Genopedia

↓

**Pairwise similarity** between Diseases or Genes

↓

**Inference Methods**

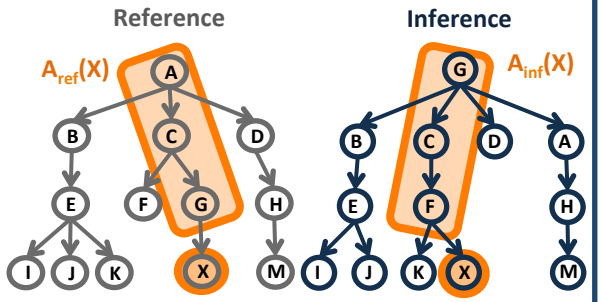
1. Our new method: Parent Promotion
2. Classical: MWST (Minimum Weight Spanning Tree)
3. Ideker Lab: CliXO[2] (Clique Extracted Ontology)

↓

**Evaluation**

**1. Edge Correctness (EC):**
What **% of edges** in **reference ontology**[3] are preserved in **inferred ontology**

**2. Ancestor Correctness (AC):**
What **% of ancestors** in **inferred ontology** are preserved in **reference ontology**




***AC(X)** = Jaccard($A_{ref}(X)$,$A_{inf}(X)$) = **2/4 = 0.5***

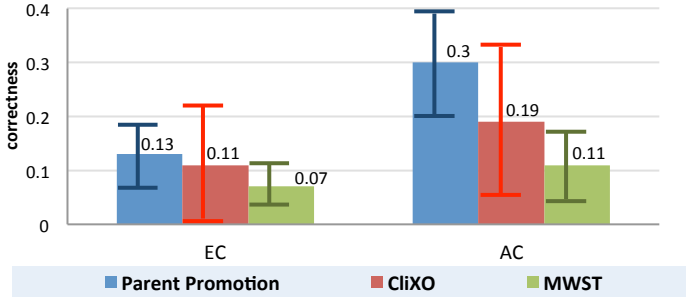
**References**

1. Desmond-Hellmann et al., *Toward precision medicine: Building a knowledge network for biomedical research and a new taxonomy of disease*, National Research Council, 2011
2. Kramer et al., *Inferring gene ontologies from pairwise similarity data*, Bioinformatics, 2014
3. Rogers, F. B. *Medical subject headings.* Bulletin of the Med. Libr. Assoc., 1963

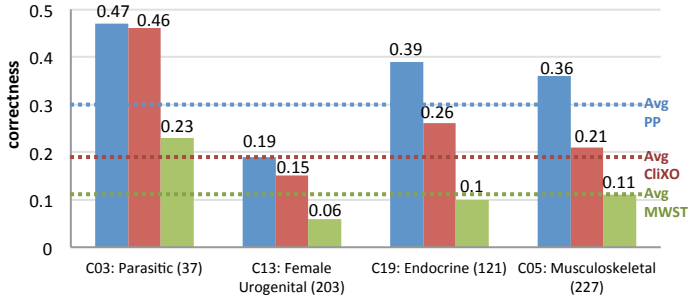
**Acknowledgement**

We thank Michael Kramer for providing the implementation of CliXO along with thoughtful advice, members of Tufts Bioinformatics and Computational Biology group for helpful feedback and comments. We gratefully acknowledge the support of NIH R01 HD076140.

## Results (Averaged across MeSH disease trees)




## Observations (AC for some MeSH disease trees)




## Discussion

1. What is the right reference? In the absence of ground truth we are limited to comparing to existing taxonomies. However, comparing our results to MeSH does tell us about the degree of molecular influence in MeSH disease trees.
2. Overall, performance is correlated across methods; some trees are more consistent with molecular data than others.
3. Performance is better on MeSH trees with fewer nodes.
4. Some large trees with well-studied complex disorders have higher scores too. (e.g., C19: diabetes and ovarian cancer; C05: arthritis).
5. No penalties for false positives. Other evaluation methods?

## Future Work

1. Parent Promotion: improve metric for deciding whom to promote
2. Experiment with other machine learning algorithms for inference
3. Incorporation of other molecular information for similarity measurement
4. Problem of "gold standard": try other reference ontology (i.e., SNOMED-CT) or build a small true disease ontology

Ultimately, our goal is to combine **medical ontologies** with **molecular information**.