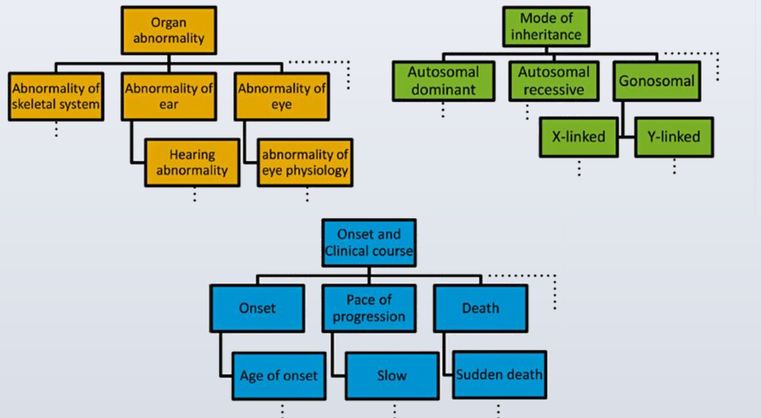
Evaluating the Effectiveness of Using Literature Features for Automated Protein Phenotype Prediction using PHENOstruct

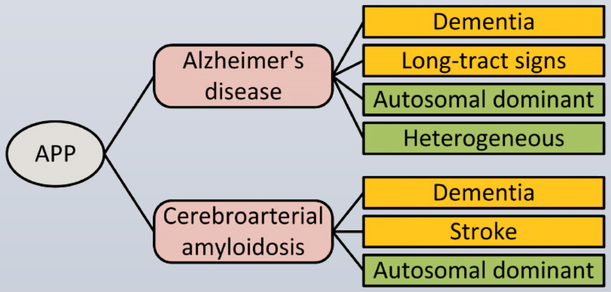
Morteza Pourreza Shahri, Indika Kahanda
Gianforte School of Computing, Montana State University

Human Phenotype Ontology (HPO)

- Recently developed ontology (Robinson et al. 2008)
- Describes disease-related phenotypic abnormalities in human
- Based on knowledge bases of human genes and genetic disorders (e.g. OMIM)
- Three independent namespaces: Organ abnormality, Inheritance, Onset

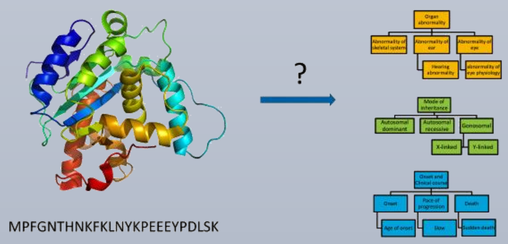


- E.g. gene-disease-HPO annotations for APP gene



The HPO prediction problem

- Only a small fraction of human protein coding genes with HPO annotations (~3000)
- Experimental determination of HPO categories for human proteins is a highly-resource-consuming task
- One of the prediction tasks in CAFA2
- PHENOstruct is the first computational method for HPO term prediction



PHENOstruct

- Uses Structured SVM
- Can capture information from the inter-relationships between the labels
- Has the advantage of not having to train multiple classifiers
- Predicted labels are hierarchically consistent

Data

Labels

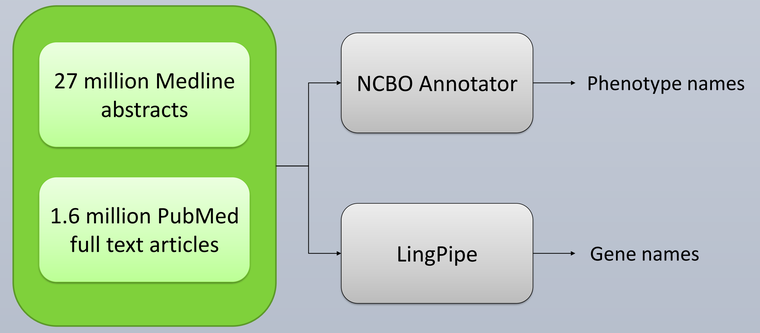
- Genes-to-HPO annotations extracted from HPO website
- HPO terms annotated to less than 10 proteins removed

	GENES	TERMS
Organ	3407	2475
Inheritance	3049	12
Onset	1053	16

Features

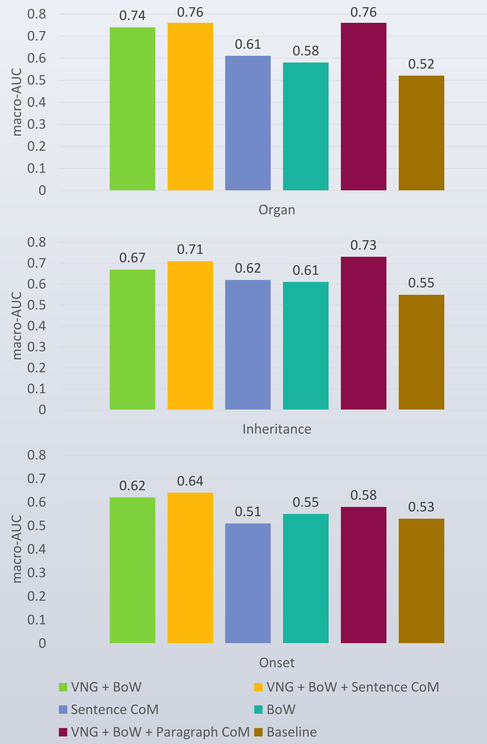
- **Variants**: All the disease variants in the human genome and their associated diseases from UniProt
- **Network**: PPI and other functional associations data (co-expression, co-occurrence, etc.) from BioGRID, STRING, and GeneMANIA
- **GO**: Experimentally derived annotations

- **Literature**: Asbtracts and full text articles from the biomedical literature:
  1. **Simple BoW:** A bag-of-words (BoW) representation for each protein using same-sentence word occurrences
  2. **Co-Mentions**: Co-mention (CoM) features in which features are the HPO categories themselves and the feature values are the frequencies of each HPO term occurring within sentences that contain protein names

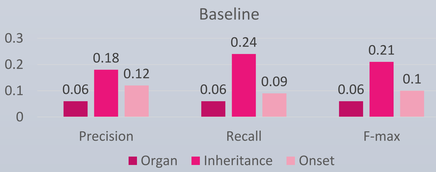


Experimental results

- We use macro-AUROC (Area Under the Receiver Operating Characteristic curve) values obtained through five fold cross validation
- VNG stands for Variants + Network + GO



- Baseline method, using CoM features as predictions themselevs (no learning)



Conclusion and future work

- Addition of literature features, specially the novel CoM features, to the original set of features improves performance
- Using PHENOstruct produces better results in comparison with the baseline method
- Paragraph-level CoM provides better performance for the Organ and Inheritance subontologies while the sentence-level CoM improves performance for the Onset subontology
- Best performance is seen in the Organ subontology using BoW and paragraph-level CoM
- Low precision and high recall of the baseline show that the literature features are able to capture relevant information, and false positive rate is high because we are able to identify many different comentions
- We plan to explore the effectiveness of using different combinations of sentence-level and paragraph-level CoM, developing a comention filter, incorporating larger spans of text, e.g. document-level, for CoM and BoW.

References

Kahanda I, Funk C, Verspoor K and Ben-Hur A. PHENOstruct: Prediction of human phenotype ontology terms using heterogeneous data sources [version 1; referees: 2 approved]. F1000Research 2015, 4:259 (doi: 10.12688/f1000research.6670.1)