Contribution of features based on sequence, predicted PPIs and GO similarities to the prediction of gene-HPO associations

Branislava Gemović (gemovic@vin.bg.ac.rs), Radoslav Davidović, Neven Šumonja, Nevena Veljković and Vladimir Perović

Center for Multidisciplinary Research, Institute of Nuclear Sciences Vinča, University of Belgrade, Serbia

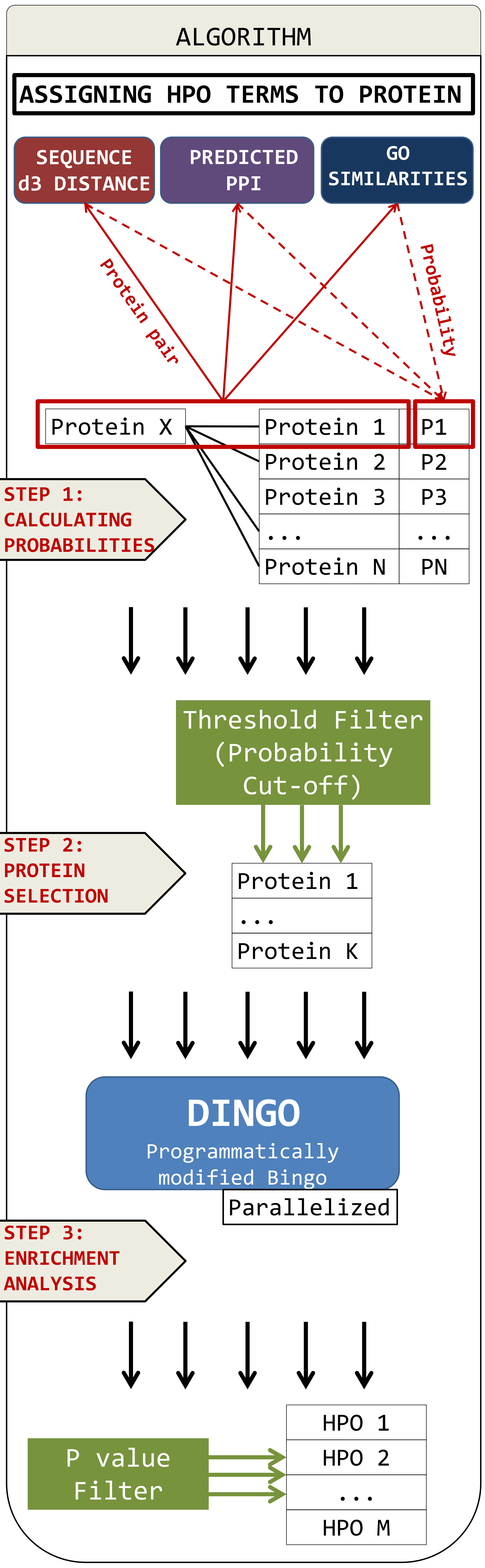
INTRODUCTION

Human Phenotype Ontology (HPO) organizes human diseases and associated genes into hierarchical classes based on the phenotypes they present. Although predicting gene-HPO relationships has important role in disease gene prioritization, this area of research is, so far, poorly investigated. Critical Assessment of protein Function Annotation (CAFA) challenge is boosting this type of research. Although there are many suggested features that can be used for HPO prediction, recent study implied that different predictors have varying contributions to the prediction performance.

AIM

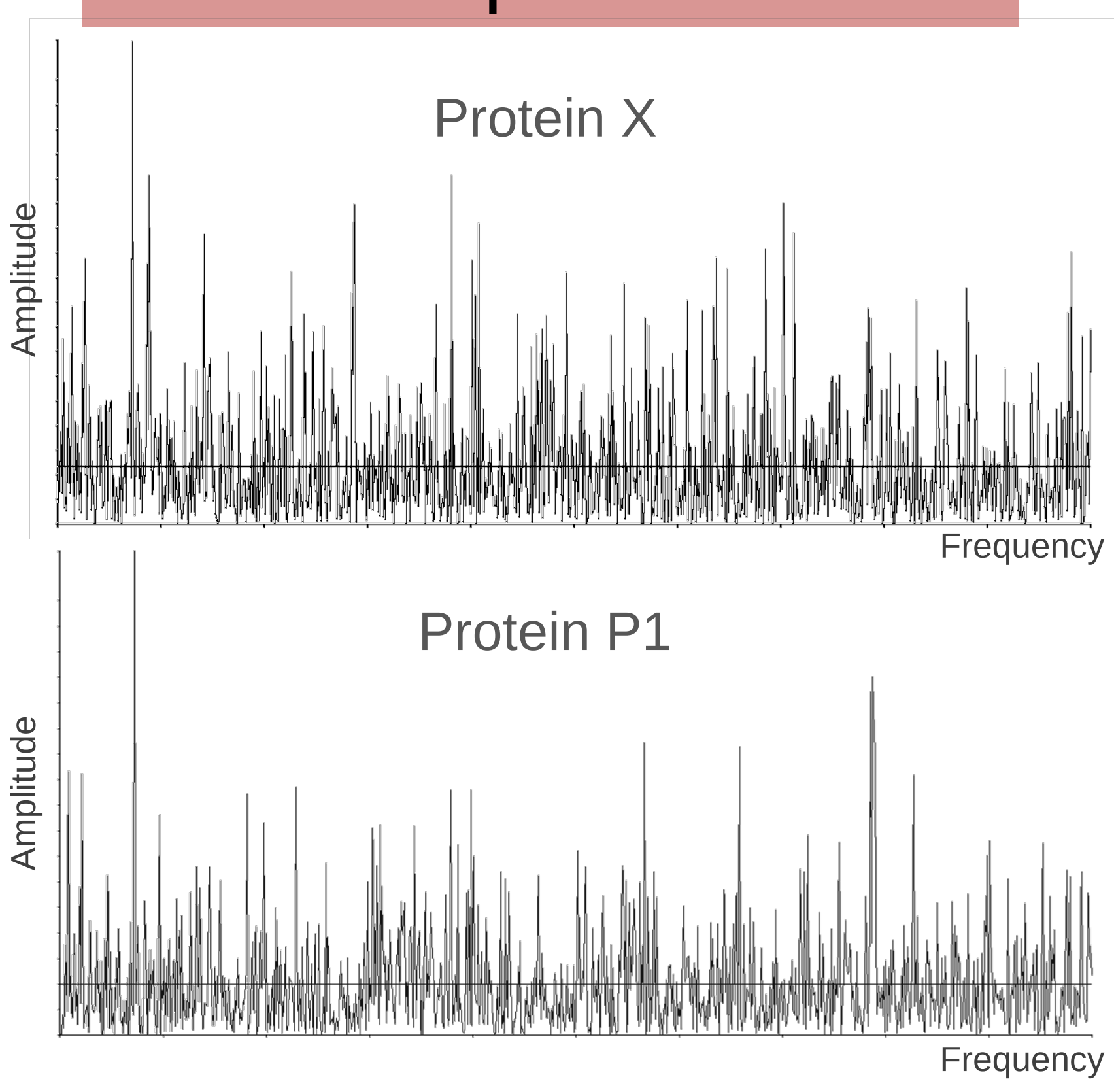
To determine common and unique landscapes of gene-HPO associations predicted by different predictors, we compered methods based on sequence, predicted protein-protein interactions (PPIs) and GO similarities.

ALGORITHM



FEATURES AND DATASET

Sequence



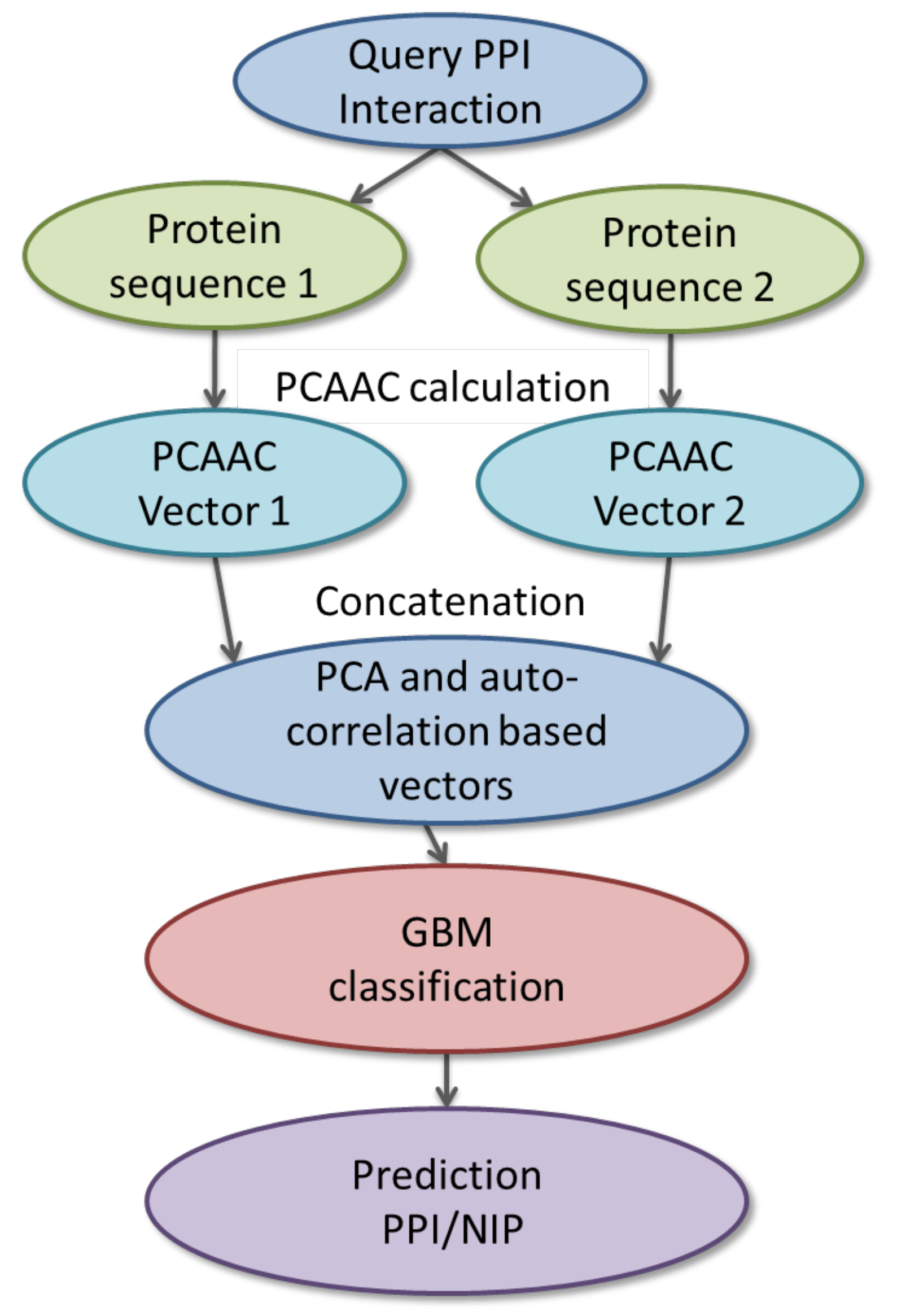
DIFFERENCE BETWEEN SEQUENCES SPECTRAL REPRESENATATIONS

$$d3(X_1, X_2) = \frac{1}{N}\sum_{n=1}^{N/2}|S_1(n) - S_2(n)|$$

where S_1 and S_2 are informational spectra of sequences X_1 and X_2 and N is the resolution of the spectrum

Predicted PPIs

PCAAC ALGORITHM



GO similarities

WANG'S METHOD

$$\mathrm{Sim}(G_1, G_2) = \frac{\sum_{1 \le i \le m} \mathrm{Sim}(go_{1i}, GO_2) + \sum_{1 \le j \le n} \mathrm{Sim}(go_{2j}, GO_1)}{m+n}$$

computing semantic similarity using the topology of the GO graph structure

IMPLEMENTATION IN R GOSemSim package

Dataset

HPO version 2016-09-03

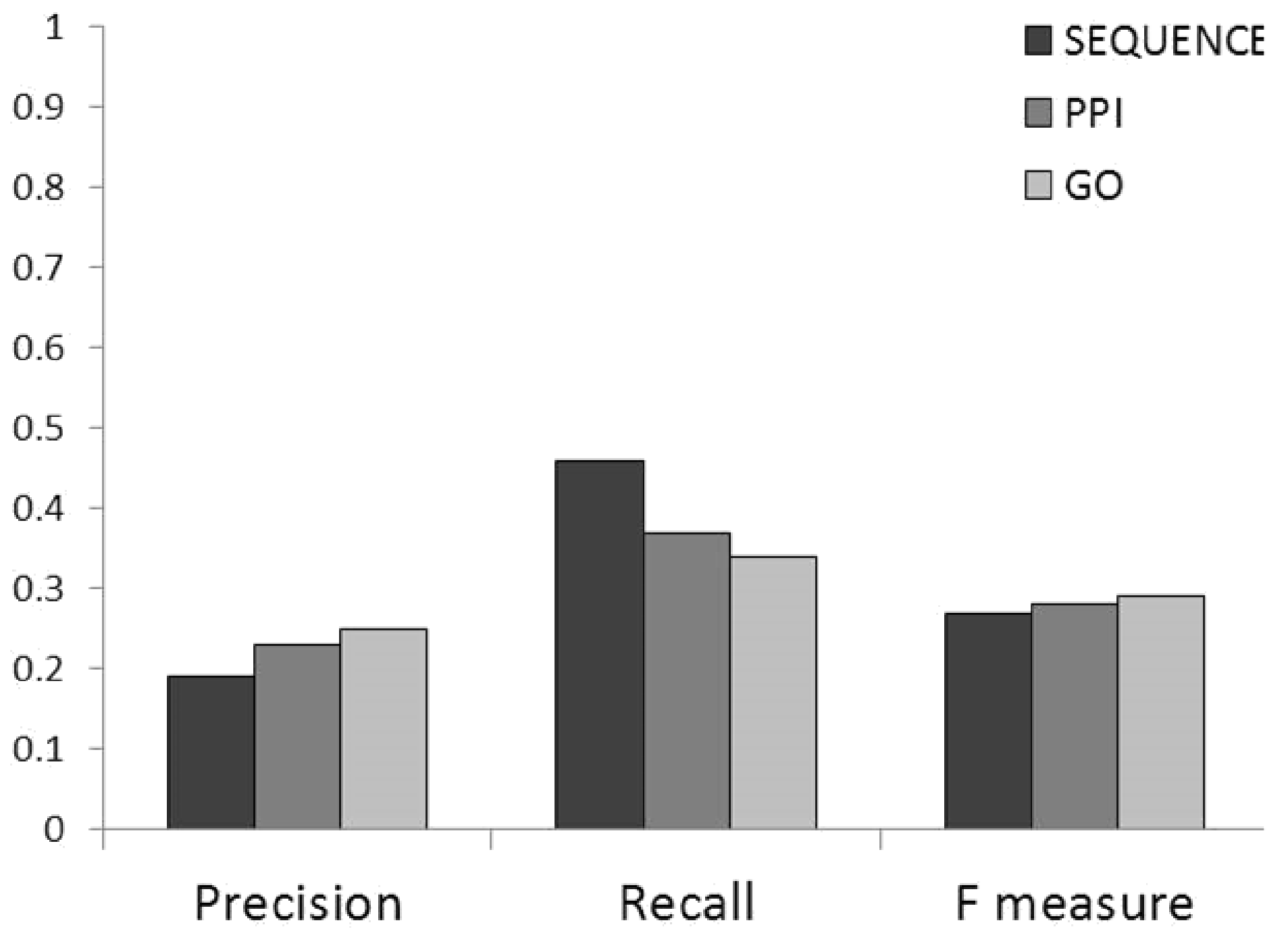
ONTOLOGY:
phenotype_annotation.tab

GENE-HPO ANNOTATIONS:
genes_to_phenotype.txt

3051 annotated genes
~44 HPO terms per gene

RESULTS

Results 1: Performance of different methods in predicting gene-HPO associations

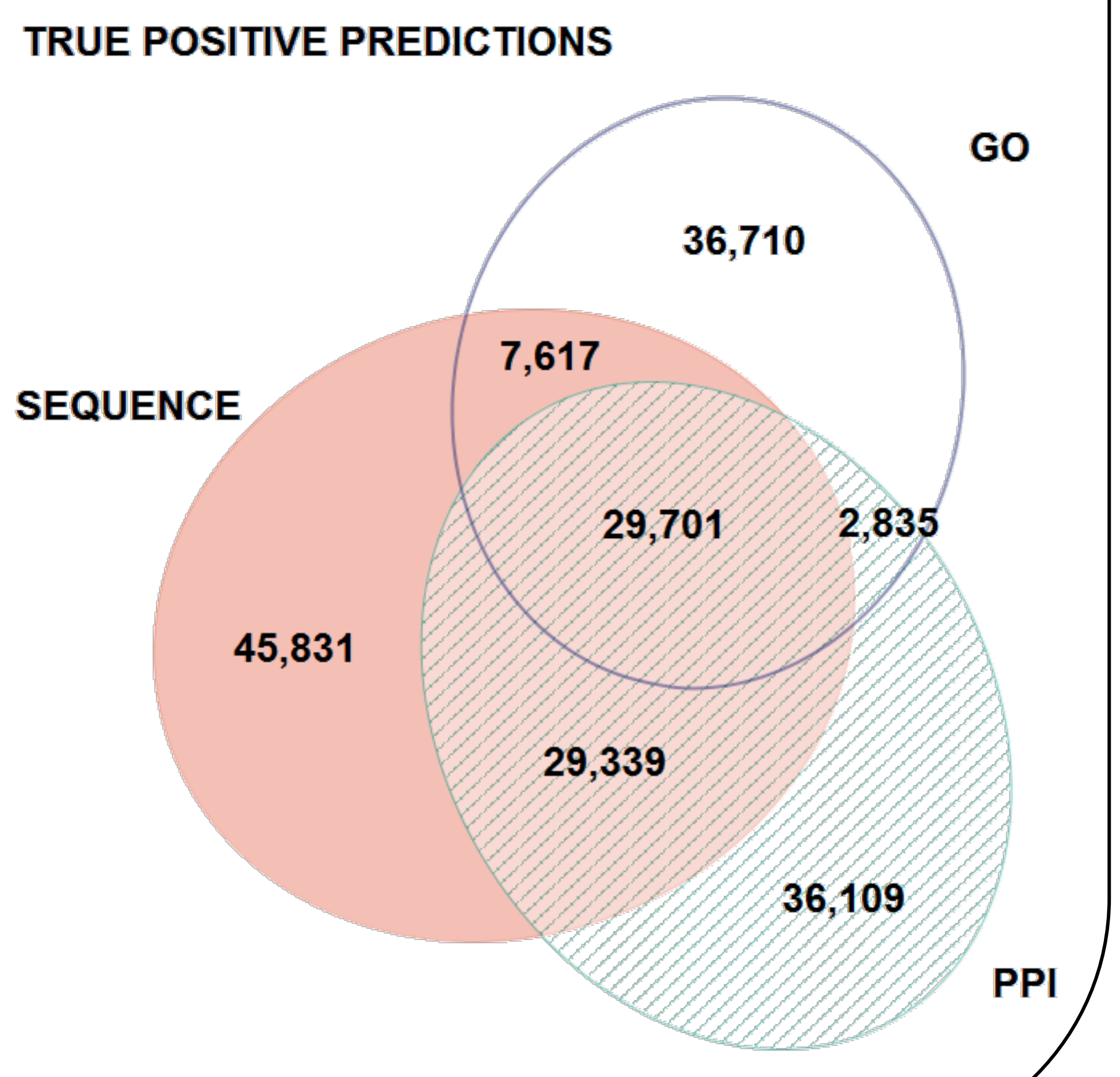


	TP	FP	FN
SEQUENCE			
PREDICTED PPIs			
GO SIMILARITIES			

Results 2: Contributions of analysed features to predicted gene-HPO associations

METHOD	
SEQUENCE d3 DISTANCE	
PREDICTED PPIs	
GO SIMILARITIES	
SEQUENCE d3 DISTANCE & PREDICTED PPIs	
SEQUENCE d3 DISTANCE & GO SIMILARITIES	
PREDICTED PPIs & GO SIMILARITIES	
SEQUENCE d3 DISTANCE & PREDICTED PPIs & GO SIMILARITIES	

Results 3: Common and unique contributions of analysed features to correctly predicted (TP) HPO terms



TAKE HOME MASSAGE

GO similarities are complementary to both sequence- and PPI-based models

CONCLUSIONS

- Sequence-based method, predicted PPIs and GO similarities perform similarly in predicting associations between genes and HPO terms.
- Although F measure varies slightly (0.27-0.29), there are important differences in precision and recall between sequence-based and other two methods.
- Out of 1.2M predicted terms, approximately 120 thousands were commonly predicted by all of the predictors.
- Methods based on sequence and predicted PPIs shared notably higher number of predictions compared to the number of predictions shared with the method based on the GO similarities.
- This trend is conspicuous when focusing on true positive (TP) predictions.

REFERENCES

1. Köhler S. et al. 2017. The Human Phenotype Ontology in 2017. *Nucleic Acids Research* 45:D865-D876.
2. Jiang Y. et al. 2016. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. *Genome Biology* 17:184.
3. Kahanda I., Funk C., Verspoor K. and Ben-Hur A. 2015. PHENOstruct: Prediction of human phenotype ontology terms using heterogeneous data sources. *F1000Research* 4:259.
4. Veljkovic N., Glisic S., Prljic J., Perovic V., Botta M. and Veljkovic V. 2008. Discovery of new therapeutic targets by the informational spectrum method. *Current Protein and Peptide Science* 9:493-506.
5. Sumonja N., Veljkovic N., Glisic S. and Perovic V. 2016. Protein-protein interaction prediction method based on principle component analysis of amino acid physicochemical properties. *Proceedings of Belgrade Bioinformatics Conference 2016* pp.131.
6. Yu G., Li F., Qin Y., Bo X., Wu Y. and Wang S. 2010. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* 26:976-8.
7. Maere S., Heymans K. and Kuiper M. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21:3448-9.