

DeepGOZero: Improving protein function prediction from sequence and zero-shot learning based on ontology axioms

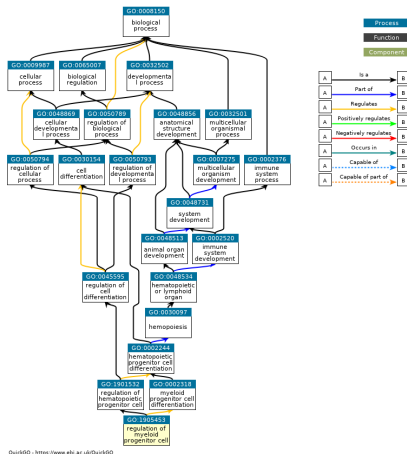
Maxat Kulmanov and Robert Hoehndorf

Bio-Ontology Research Group, Computational Bioscience
Research Center, KAUST, Thuwal, Saudi Arabia

July 14, 2022

Introduction - Gene Ontology

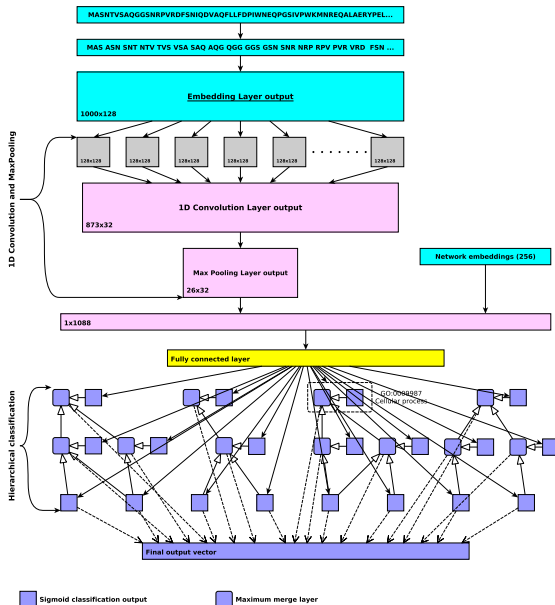
- Protein Functions - Gene Ontology (GO) annotations
- GO provides a vocabulary for describing gene products
 - Biological Process - processes to which gene or gene product contributes
 - Molecular Function - biochemical activity of a gene product
 - Cellular Component - location of a gene product where it is active



QuickGO - <https://www.ebi.ac.uk/QuickGO>

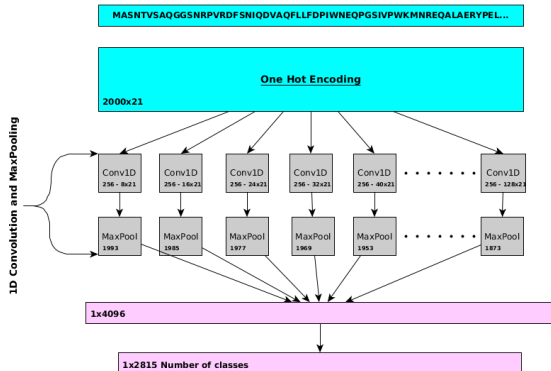
- How to incorporate the knowledge in GO into prediction model ?

Previous models - DeepGO



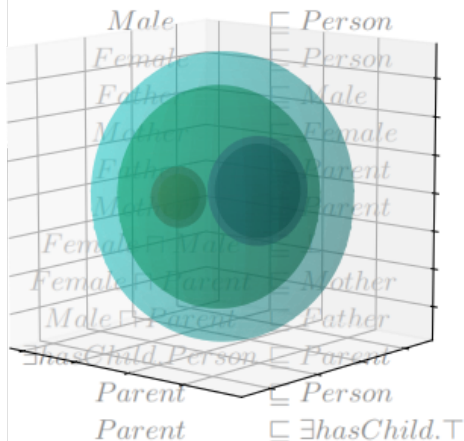
- Predicts limited number of functions
- Slow
- Low performance

Previous models - DeepGOPlus



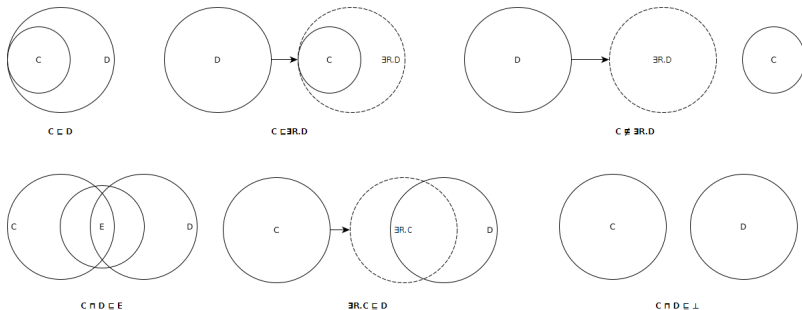
- Can predict all functions
- Fast
- Post-prediction use the ontology structure

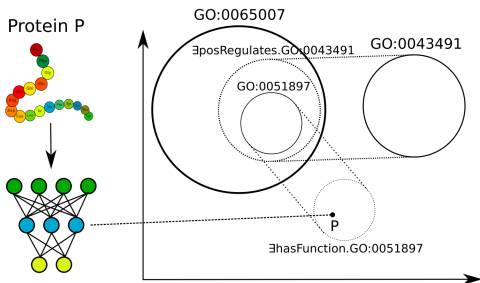
Previous models - ELEmbeddings



- Embeds ontology semantics into \mathbb{R}^n
- Directly learns from axioms
- Applied to protein-protein interaction prediction

Previous models - ELEmbeddings





- Embeds proteins into ELEmbeddings space

- Prediction score of a class c for a protein p

- $y'_c = \sigma(f_\eta(p) \cdot (f_\eta(hF) + f_\eta(c))^T + r_\eta(c))$

- Loss function

- $L = \frac{1}{N} \sum_{i=1}^N BCELoss(y_{c_i}, y'_{c_i}) + ELLoss$

$$L_{NF1} = \frac{1}{|NF1|} \sum_{c,d \in NF1} \max(0, \|f_\eta(c) - f_\eta(d)\| + r_\eta(c) - r_\eta(d) - \gamma)$$

$$L_{NF2} = \frac{1}{|NF2|} \sum_{c,d,e \in NF2} \max(0, \|f_\eta(c) - f_\eta(d)\| - r_\eta(c) - r_\eta(d) - \gamma) + \max(0, \|f_\eta(c) - f_\eta(e)\| - r_\eta(c) - \gamma) + \max(0, \|f_\eta(d) - f_\eta(e)\| - r_\eta(c) - \gamma) + \max(0, \min(r_\eta(c), r_\eta(d)) - r_\eta(e) - \gamma)$$

$$L_{NF3} = \frac{1}{|NF3|} \sum_{r,c,d \in NF3} \max(0, \|f_\eta(c) - f_\eta(r) - f_\eta(d)\| - r_\eta(c) - r_\eta(d) - \gamma)$$

$$L_{NF4} = \frac{1}{|NF4|} \sum_{c,r,d \in NF4} \max(0, \|f_\eta(c) + f_\eta(r) - f_\eta(d)\| + r_\eta(c) - r_\eta(d) - \gamma)$$

- Our data - 50% identity, similarity based split (UniprotKB-Swissprot, Nov 2021)
 - 81% Training, 9% Validation, 10% Testing
- NetGO2 data - time based split
 - Training – all data annotated in December 2018 or before
 - Validation – proteins experimentally annotated from January 2019 to January 2020 and not before January 2019
 - Testing – proteins experimentally annotated between February 2020 and October 2020 and not before February 2020

Comparison Methods - our baseline

- **DiamondScore** - Sequence similarity based predictions
- **MLP** - Multi-Layer Perceptron (DeepGOZero without ELEmbeddings)
- **DeepGOCNN** - Convolutional Neural Networks based predictions
- **DeepGOPlus** - DeepGOCNN + DiamondScore

Comparison Methods - other methods

- **NetGO2** - Ensemble method, combines predictions based on sequence similarity, protein–protein interactions, literature, recurrent neural networks, GO term frequency
- **DeepGraphGO** - Graph Convolutional Neural Networks based predictions. Uses InterPRO domain annotations and protein–protein interactions
- **TALE+** - transformer based neural networks predictions combine with sequence similarity

Results : Our data - MFO

Method	F_{\max}	S_{\min}	AUPR	AUC
DiamondScore	0.623	10.145	0.380	0.747
MLP	0.657	9.857	0.655	0.882
MLP + DiamondScore	0.670	9.551	0.649	0.886
DeepGOCNN	0.430	13.601	0.393	0.765
DeepGOPlus	0.634	10.072	0.636	0.844
DeepGOZero	0.657	9.808	0.657	0.903
DeepGOZero + DiamondScore	0.668	9.595	0.673	0.906

Results : Our data - BPO

Method	F_{\max}	S_{\min}	AUPR	AUC
DiamondScore	0.444	45.040	0.313	0.610
MLP	0.460	43.987	0.435	0.793
MLP + DiamondScore	0.486	43.822	0.449	0.797
DeepGOCNN	0.344	48.543	0.289	0.672
DeepGOPlus	0.462	44.485	0.421	0.726
DeepGOZero	0.451	44.621	0.422	0.798
DeepGOZero + DiamondScore	0.482	44.058	0.446	0.803

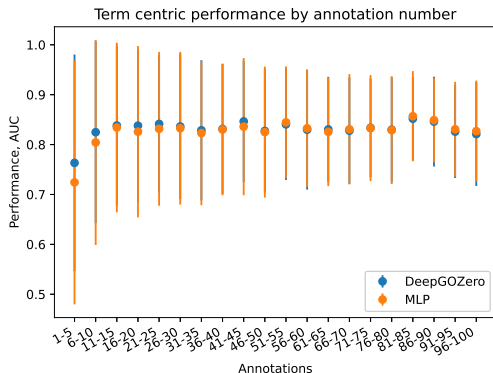
Results : Our data - CCO

Method	F_{\max}	S_{\min}	AUPR	AUC
DiamondScore	0.581	11.092	0.352	0.648
MLP	0.667	10.523	0.670	0.846
MLP + DiamondScore	0.666	10.526	0.654	0.851
DeepGOCNN	0.641	11.396	0.645	0.775
DeepGOPlus	0.672	10.591	0.667	0.821
DeepGOZero	0.661	10.681	0.665	0.854
DeepGOZero + DiamondScore	0.667	10.615	0.701	0.860

Results : Time based split

Method	F_{\max}			S_{\min}			AUPR			AUC		
	MFO	BPO	CCO	MFO	BPO	CCO	MFO	BPO	CCO	MFO	BPO	CCO
DiamondScore	0.627	0.407	0.625	5.503	25.918	9.351	0.427	0.272	0.412	0.836	0.643	0.682
DeepGOCNN	0.589	0.337	0.624	6.417	27.235	10.617	0.565	0.271	0.623	0.867	0.694	0.834
DeepGOPlus	0.661	0.419	0.655	5.407	25.603	9.374	0.667	0.342	0.663	0.913	0.737	0.869
MLP	0.667	0.419	0.656	5.326	24.825	9.688	0.672	0.359	0.650	0.921	0.738	0.839
MLP + DiamondScore	0.659	0.446	0.662	5.316	24.904	9.545	0.664	0.364	0.651	0.924	0.740	0.846
DeepGOZero	0.662	0.396	0.662	5.322	25.838	9.834	0.668	0.337	0.645	0.930	0.717	0.809
DeepGOZero + DiamondScore	0.655	0.432	0.675	5.337	25.439	9.391	0.665	0.356	0.654	0.938	0.725	0.827
NetGO2 (Web-server)	0.698	0.431	0.662	5.187	25.076	9.473	0.701	0.343	0.627	0.856	0.635	0.772
DeepGraphGO	0.671	0.418	0.679	5.374	25.866	9.165	0.647	0.364	0.669	0.930	0.815	0.857
TALE+	0.466	0.382	0.661	8.136	26.308	9.599	0.441	0.310	0.681	0.753	0.608	0.778

Results - Specific terms



- Improves performance of specific terms (< 50 annotations)

Zero shot predictions

- Prediction score of a class c for a protein p
 - $y'_c = \sigma(f_\eta(p) \cdot (f_\eta(hF) + f_\eta(c))^T + r_\eta(c))$
- Definition axioms
 - *serine-type endopeptidase inhibitor activity* (GO:0004867) **is equivalent to** *molecular function regulator* (GO:0098772) and *negatively regulates* (RO:0002212) some *serine-type endopeptidase activity* (GO:0004252)

Zero shot predictions

Ontology	Term	Name	AUC (test)	AUC (all)	AUC (trained)	AUC (trained mlp)
mf	GO :0001227	DNA-binding transcription repressor activity, RNA polymerase II-specific	0.257	0.405	0.932	0.926
mf	GO :0001228	DNA-binding transcription activator activity, RNA polymerase II-specific	0.574	0.699	0.948	0.944
mf	GO :0003735	structural constituent of ribosome	0.400	0.194	0.940	0.942
mf	GO :0004867	serine-type endopeptidase inhibitor activity	0.972	0.967	0.985	0.984
mf	GO :0005096	GTPase activator activity	0.847	0.870	0.938	0.960
bp	GO :0000381	regulation of alternative mRNA splicing, via spliceosome	0.855	0.865	0.906	0.886
bp	GO :0032729	positive regulation of interferon-gamma production	0.870	0.919	0.932	0.906
bp	GO :0032755	positive regulation of interleukin-6 production	0.719	0.819	0.884	0.873
bp	GO :0032760	positive regulation of tumor necrosis factor production	0.861	0.906	0.925	0.867
bp	GO :0046330	positive regulation of JNK cascade	0.855	0.894	0.904	0.916
bp	GO :0051897	positive regulation of protein kinase B signaling	0.772	0.864	0.888	0.915
bp	GO :0120162	positive regulation of cold-induced thermogenesis	0.637	0.789	0.738	0.835
cc	GO :0005762	mitochondrial large ribosomal subunit	0.889	0.975	0.874	0.916
cc	GO :0022625	cytosolic large ribosomal subunit	0.898	0.969	0.893	0.849
cc	GO :0042788	polysomal ribosome	0.858	0.950	0.889	0.780
cc	GO :1904813	ficolin-1-rich granule lumen	0.653	0.782	0.792	0.900
		Average	0.745	0.804	0.898	0.900

Zero shot predictions

Ontology	All classes		Defined classes	
	Num. classes	AUC	Num. classes	AUC
MFO	4,791	0.804	95	0.862
BPO	11,092	0.737	4,598	0.786
CCO	1,492	0.819	151	0.915

DeepGOZero :

- Achieves best performance in class-centric evaluations
- Improves predictions for specific classes with few annotations
- Can make zero shot predictions

- Limitations
 - Uses only InterPRO domain annotations
 - Limitations of ELEmbeddings : class intersection, many-to-many relations
 - Does not use all types of axioms
- Future work
 - Incorporate sequence, interactions, structure, literature
 - EL Box Embeddings

Thanks !