MetaGOPlus: Improving Gene Ontology Prediction of Proteins Using Deep Residual Network with Hierarchical Classification

Yiheng Zhu [1, 2] , Chengxin Zhang [2] , Rucheng Diao [2] , Xiaogen Zhou [2] ,

Peter L. Freddolino [2] , Dongjun Yu [1] , Yang Zhang [2]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, 200 Xiaolingwei, Nanjing, Jiangsu, China

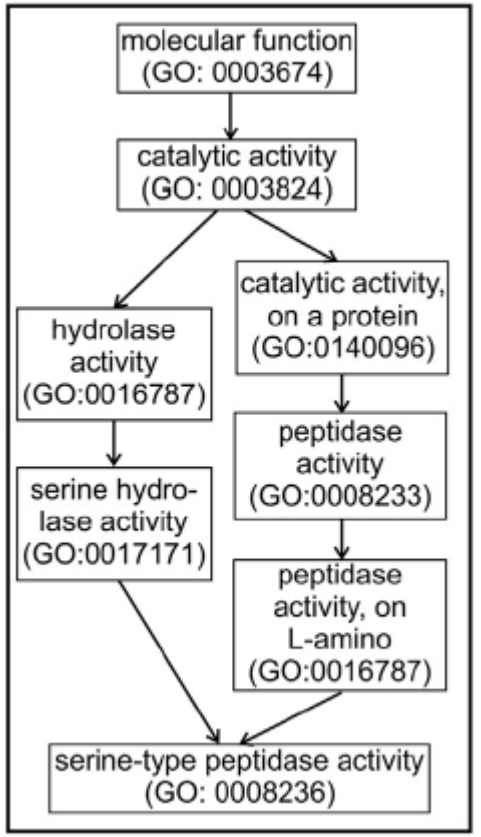
[2] Department of Computational Medicine and Bioinformatics, University of Michigan, 100 Washtenaw Avenue, Ann Arbor, Michigan, USA

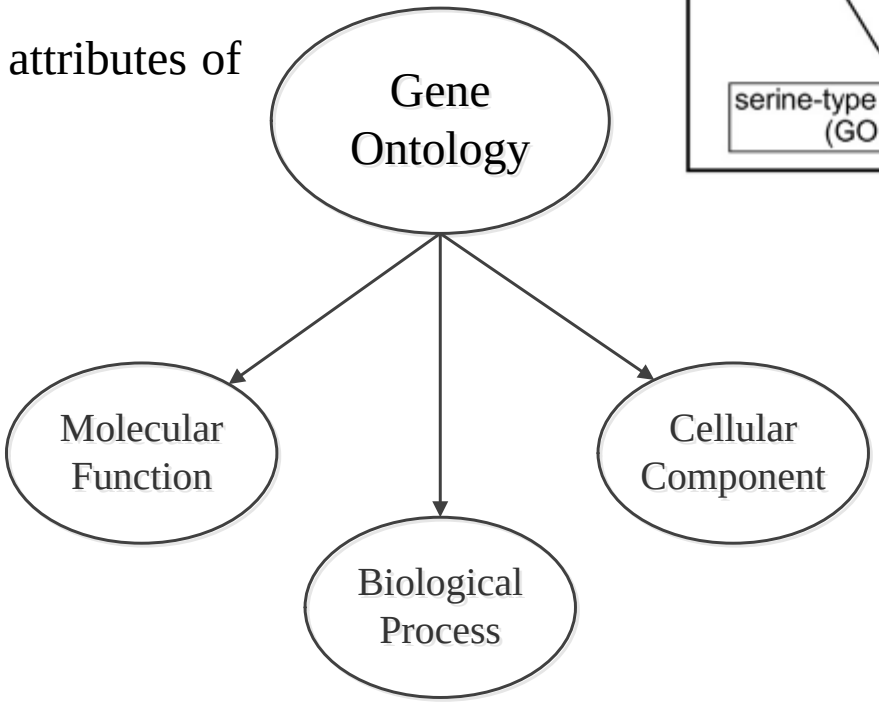
Outline

- Background and Motivation
- Benchmark Dataset and Method
- Computational Experiment Result
- Conclusion and Future Work

Background and Motivation

- Gene Ontology (GO) is a bioinformatics term to represent gene and gene product attributes across all species.
- GO has been widely used to annotate functions of proteins.
- Accurate identification of GO attributes from proteins provides critical help in understanding the biological activities of proteins.
- We proposed a new pipeline, MetaGOPlus, to predict the GO attributes of proteins.





Benchmark Datasets

Training Dataset (Train_121657): 121657 proteins (30 < length < 1000)

Evaluation Dataset (Eval_1604): 1604 proteins (30 < length < 1000)

Test Dataset (Test_1464): 1464 proteins (30 < length < 1000)

Benchmark dataset	Number of MF proteins	Number of MF terms	Number of BP proteins	Number of BP terms	Number of CC proteins	Number of CC terms
Training dataset	75746	5153	79728	15052	85719	2227
Evaluation dataset	788	995	1269	4414	923	478
Test dataset	699	906	1163	4359	878	445

Note : (1) In Train_121657, the GO terms whose frequency are less than 15, 50, and 15 are removed. So, the numbers of GO terms are 1717, 2617, and 1010, for MF, BP, and CC proteins.

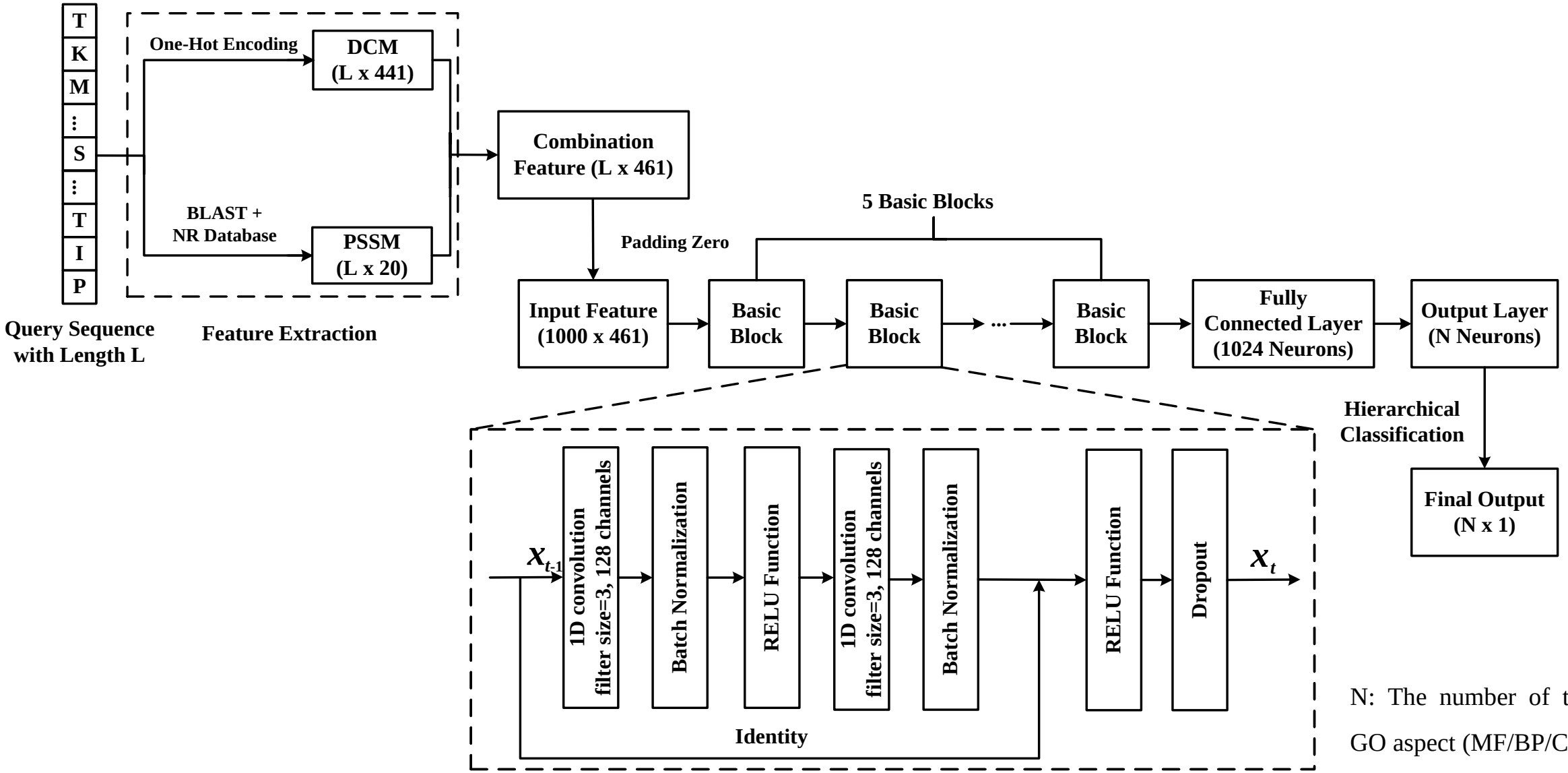
(2) The identity is less than 30% for any two sequences between test dataset and training dataset or test dataset and evaluation dataset.

Method

- **Sequence and sequence-profile based pipeline [1]**
- **Protein-protein interaction based pipeline [1]**
- **Naive based pipeline [1]**
- **Deep learning based pipeline**

[1] Chengxin Zhang, Wei Zheng, Peter L Freddolino, Yang Zhang. MetaGO: Predicting Gene Ontology of non-homologous proteins through low-resolution protein structure prediction and protein–protein network mapping [J]. Journal of molecular biology, 2018, 430: 2256-2265.

Deep Learning based Pipeline



N: The number of terms for GO aspect (MF/BP/CC)

The workflow of deep learning based pipeline

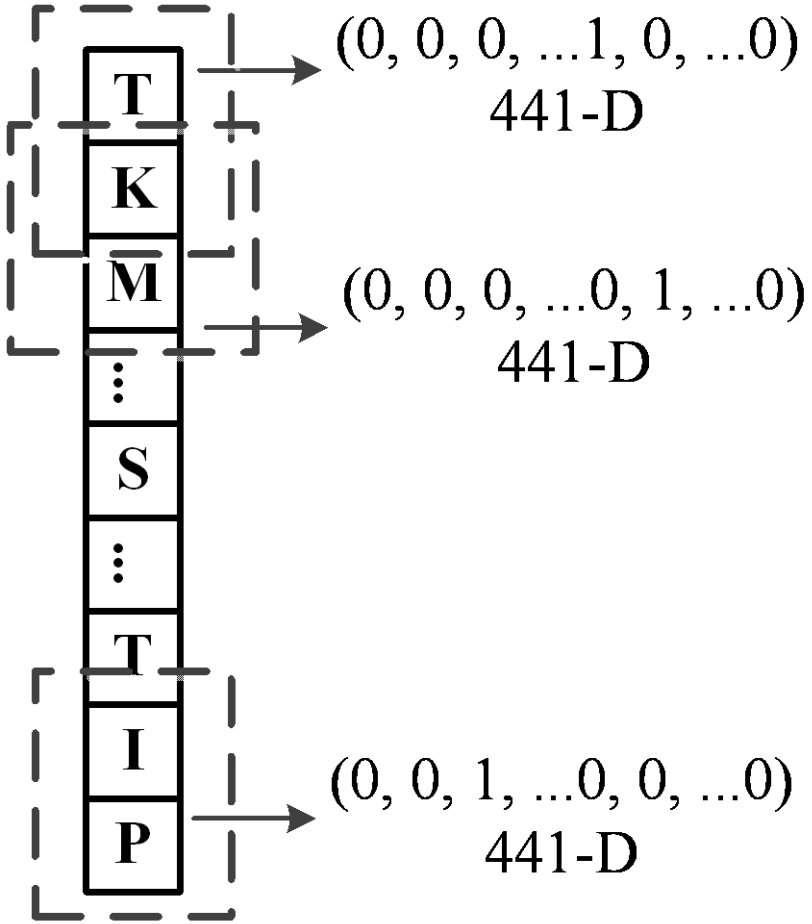
Feature Extraction

- Position Specific Scoring Matrix (PSSM)

Given a protein with L residues, we use BLAST software to search against NR database with three iterations to generate its PSSM with L rows and 20 columns.

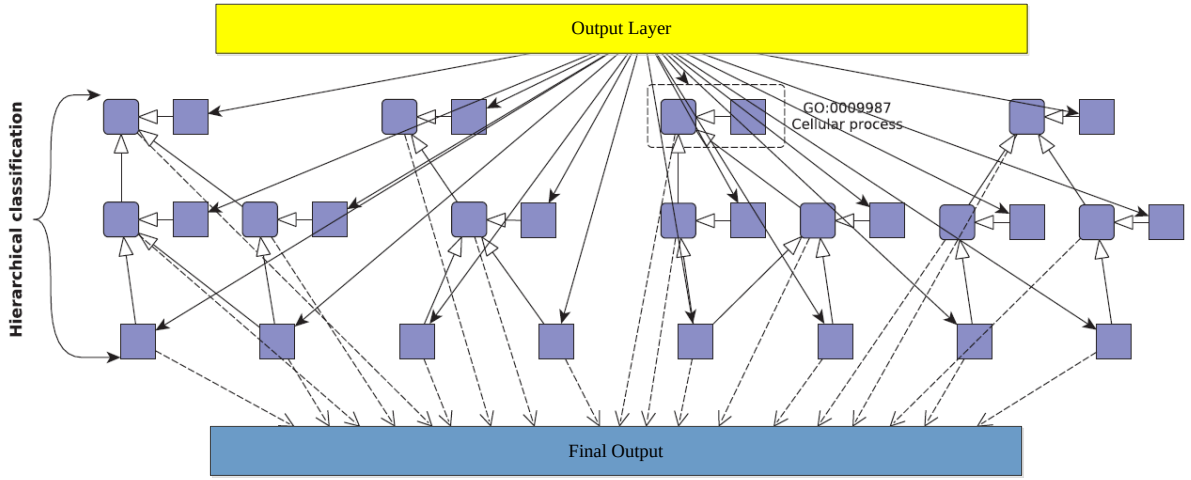
- Dipeptide Composition Matrix (DCM)

In a protein, there are 21 types of amino acids at most (20 common amino acids and 1 other amino acid, such as ‘B’, ‘U’, and ’X’). Therefore, there are 441 (21 x 21) possible amino acid pairs (A_1A_1, …, A_1A_{21}, …, $A_{21}A_1$, …, $A_{21}A_{21}$). Given a protein with L residues, its DCM can be represented as a matrix with L rows and 441 columns by one-hot coding (The last row of DCM is padded with 0).



Hierarchical Classification [2]

In output layer, each neuron can be viewed as a GO term (GO_i), the corresponding output value (P_i) can be viewed as the probability that a protein has term GO_i. Moreover, the children terms of GO_i in output layer are denoted as $GO_{i,1}$, $GO_{i,2}$, …$GO_{i,j}$, and their outputs are denoted as $P_{i,1}$, $P_{i,2}$, …$P_{i,j}$. In hierarchical classification, the output value of GO_i should be re-calculated as follows: $P_i = \max\{P_{i,1}, P_{i,2}, \ldots P_{i,j}\}$.



[2] Kulmanov, Maxat, Mohammed Asif Khan, and Robert Hoehndorf. "DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier." *Bioinformatics* 34.4 (2018): 660-668.

Computational Experiment Result

The performances of different pipelines on Test_1464

GO aspect	Pipeline	Precision	Recall	F-max	AUPR
MF	SSPP	0.651	0.593	0.621	0.542
	PPIP	0.403	0.356	0.378	0.311
	NP	0.248	0.185	0.212	0.100
	DLP	0.565	0.434	0.491	0.416
	SPN	0.607	0.641	0.623	0.608
	SPND	0.637	0.630	0.634	0.617
BP	SSPP	0.408	0.324	0.361	0.265
	PPIP	0.327	0.303	0.315	0.227
	NP	0.230	0.198	0.213	0.122
	DLP	0.300	0.270	0.284	0.205
	SPN	0.426	0.360	0.390	0.312
	SPND	0.415	0.385	0.399	0.318
CC	SSPP	0.536	0.579	0.556	0.492
	PPIP	0.639	0.468	0.541	0.551
	NP	0.590	0.473	0.525	0.394
	DLP	0.608	0.524	0.563	0.560
	SPN	0.562	0.620	0.589	0.594
	SPND	0.612	0.596	0.604	0.627

Note :

(1) SSPP: Sequence and sequence-profile based pipeline; PPIP: Protein-protein interaction based pipeline; NP: Naive based pipeline; DLP: Deep learning based pipeline.

(2) In SSPP and PPIP, we removed all of the templates which have more than 30% sequence identity with the query.

(3) SPN: SSPP+PPIP+NP (ensemble method: neural network)

(4) SPND: SSPP+PPIP+NP+DLP (ensemble method: neural network)

(5) AUPR: Area Under the Precision-Recall curve

Conclusion

- The deep-learning-based pipeline can further improve the prediction performance of gene ontology for proteins.
- Four pipelines can learn the knowledge of GO attributes from different views and be complementary with each other.

Future Work

- Extract more effective sequence-based features (such as HHblits-based feature) and combine them with the current features to form the input of deep residual network.
- Test the performance of other deep learning models (e.g. long short-term memory (LSTM) and graph convolutional network (GCN)).