DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction

2022-10-04 / JiWung Han

Department of Artificial Intelligence
Korea University

1. Abstract & Introduction

Abstract

- **Motivation**
  - Automated function prediction (AFP) of proteins
    - a large-scale multi-label classification problem
  - Two limitations of most network-based methods for AFP
    1. a single model must be trained **for each species**
    2. protein **sequence information** is totally ignored
  - Limitations cause
    - Weaker performance than sequence-based methods
    - The challenge is how to develop a powerful network-based method for AFP

Abstract

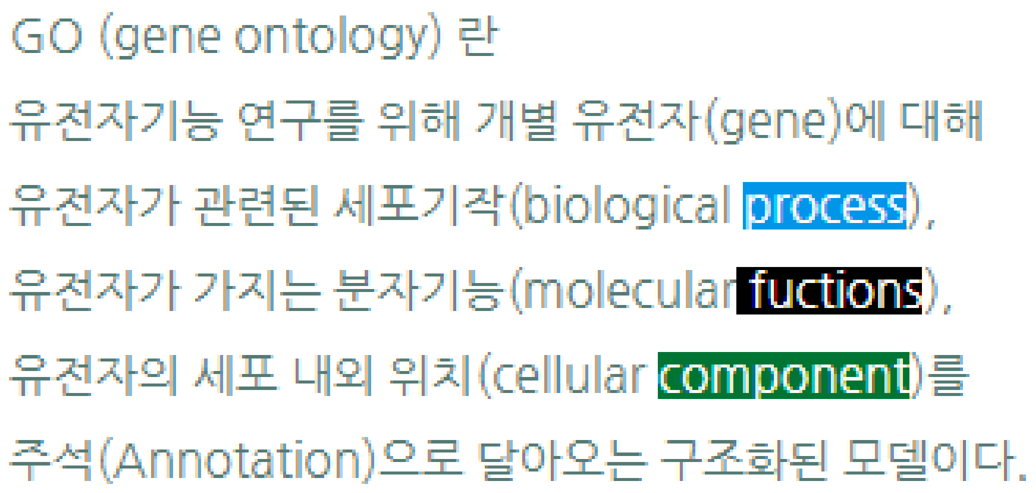
- **Results**
  - Multispecies graph neural network-based method for AFP,
    - Both protein sequence and high-order protein network information
  - Our multispecies strategy allows one single model to be trained for all species
    - Indicating a larger number of training samples than existing methods.
    - Extensive experiments with a large-scale dataset
  - Further confirm the effectiveness multispecies strategy
  - Integrate DeepGraphGO into the state of-the-art ensemble method, NetGO, as a component and achieve a further performance improvement

Introduction

- **Proteins**
  - Building blocks of life
  - Playing many crucial roles within organisms
    - Catalyzing chemical reactions,
    - Coordinating signal pathway
    - Providing structural support to cells
  - Important to identify protein/gene functions, which are now standarized by Gene Ontology (GO)

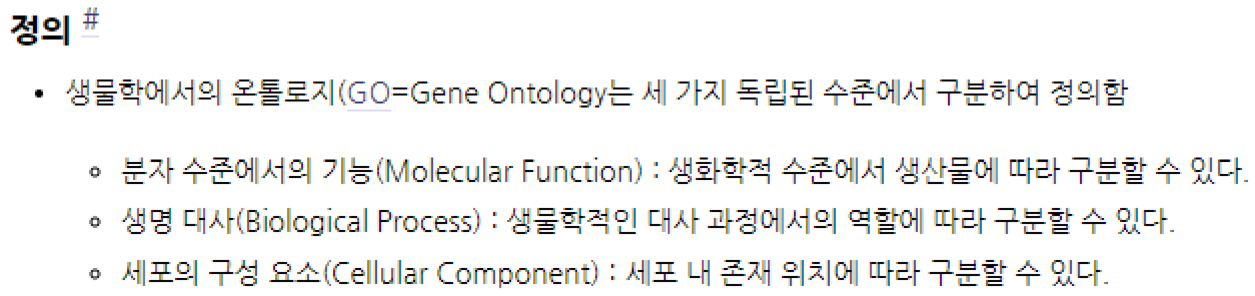
Introduction

- **Gene Ontology (GO)**
  - https://www.nature.com/articles/ng0500_25
  - Covers three biological domains: molecular function ontology (MFO), biological process ontology (BPO) and cellular component ontology (CCO)



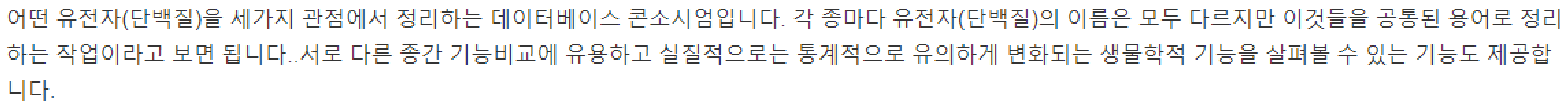
GO (gene ontology) 란

유전자기능 연구를 위해 개별 유전자(gene)에 대해
유전자가 관련된 세포기작(biological process),
유전자가 가지는 분자기능(molecular fuctions),
유전자의 세포 내외 위치(cellular component)를
주석(Annotation)으로 달아오는 구조화된 모델이다.



정의 #

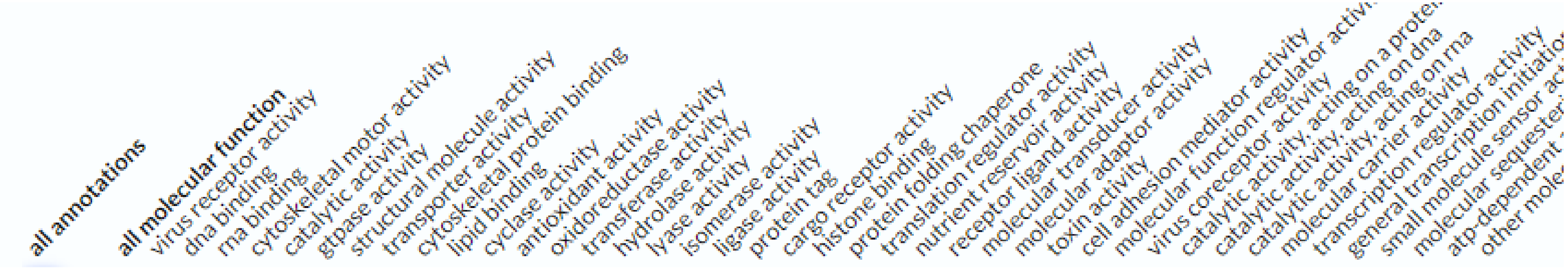
- 생물학에서의 온톨로지(GO=Gene Ontology는 세 가지 독립된 수준에서 구분하여 정의함
  - 분자 수준에서의 기능(Molecular Function) : 생화학적 수준에서 생산물에 따라 구분할 수 있다.
  - 생명 대사(Biological Process) : 생물학적인 대사 과정에서의 역할에 따라 구분할 수 있다.
  - 세포의 구성 요소(Cellular Component) : 세포 내 존재 위치에 따라 구분할 수 있다.

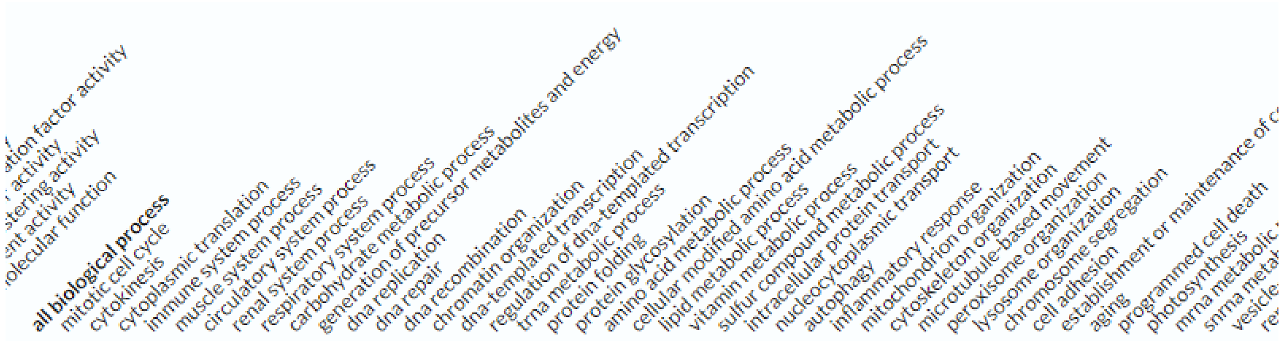


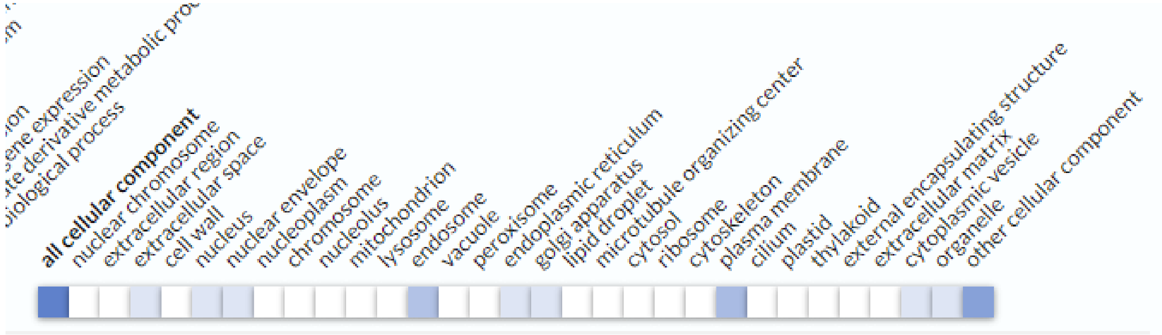
어떤 유전자(단백질)을 세가지 관점에서 정리하는 데이터베이스 콘소시엄입니다. 각 종마다 유전자(단백질)의 이름은 모두 다르지만 이것들을 공통된 용어로 정리하는 작업이라고 보면 됩니다..서로 다른 종간 기능비교에 유용하고 실질적으로는 통계적으로 유의하게 변화되는 생물학적 기능을 살펴볼 수 있는 기능도 제공합니다.

Introduction

- **P00533 · EGFR_HUMAN**
  - https://www.uniprot.org/uniprotkb/P00533/entry

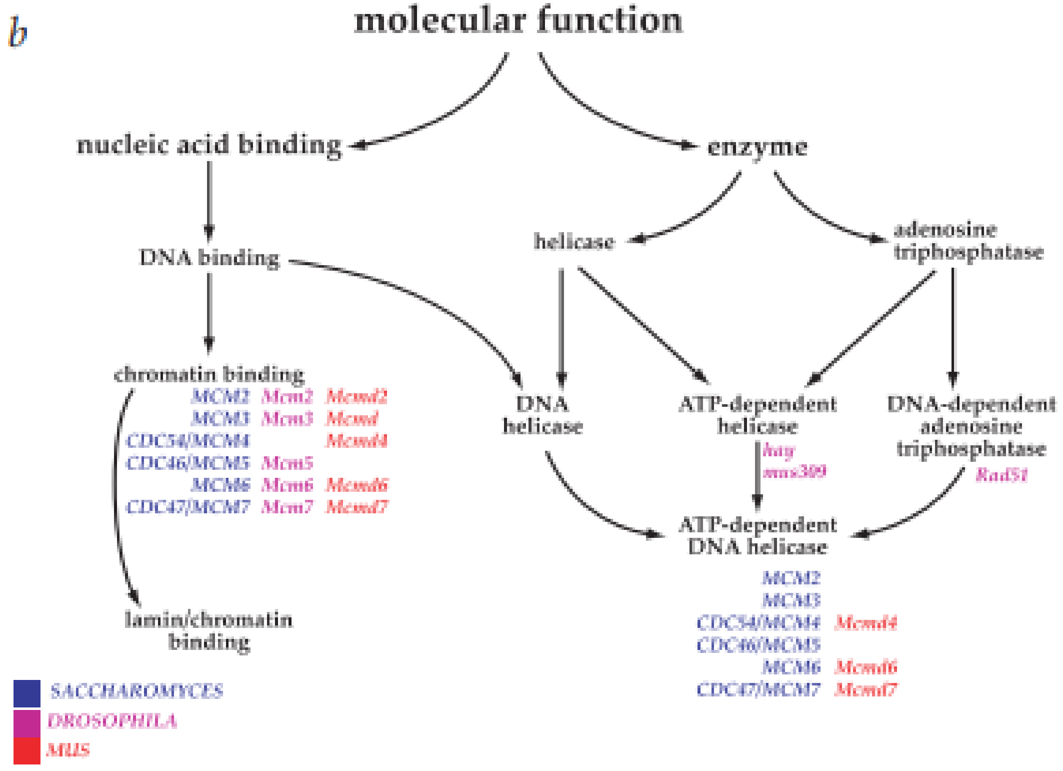

all annotations
all molecular function
virus receptor activity
dna binding
rna binding
cytoskeletal motor activity
catalytic activity
gtpase activity
structural molecule activity
transporter activity
cytoskeletal protein binding
lipid binding
cyclase activity
antioxidant activity
oxidoreductase activity
transferase activity
hydrolase activity
lyase activity
isomerase activity
ligase activity
protein tag
cargo receptor activity
histone binding
protein folding chaperone
translation regulator activity
nutrient reservoir activity
receptor ligand activity
molecular transducer activity
molecular adaptor activity
toxin activity
cell adhesion mediator activity
molecular function regulator activi
virus coreceptor activity
catalytic activity, acting on a prote
catalytic activity, acting on dna
catalytic activity, acting on rna
molecular carrier activity
transcription regulator activity
general transcription initiatio
small molecule sensor ac
molecular sequester
atp-dependent
other mole





Introduction

- **Gene Ontology (GO)**
  - Only < 0:1% proteins have experimental GO annotations due to the high cost of biochemical experiments
  - Developing an effective and efficient automatic protein function prediction (AFP) method is of great significance



Introduction

- **Critical Assessment of protein Function Annotation algorithms (CAFA)**
  - Function Special Interest Group (Function-SIG) of International Society for Computational Biology (ISCB)
    - organized a community challenge, the Critical Assessment of protein Function Annotation algorithms (CAFA)
    - CAFA1 in 2010–2011, CAFA2 in 2013–2014, CAFA3 in 2015– 2016 and CAFA4 in 2019–2020
  - CAFA3 and CAFA4,
    - provided a large number of protein sequences (around 100 000) to the participants
    - have to submit the predictions of protein functions (GO term associations)

Introduction

- **Critical Assessment of protein Function Annotation algorithms (CAFA)**
  - CAFA3 and CAFA4,
    - provided a large number of protein sequences (around 100 000) to the participants
    - have to submit the predictions of protein functions (GO term associations)
  - For building the benchmark data, then the organizers collect proteins with experimental annotations by a few months later (T1, 10 months later in CAFA3)
  - The benchmark data consists of two types of proteins
    - **no-knowledge**
      - receive at least one experimental annotation between T0 and T1
      - Without any experimental annotations before T0
    - limited-knowledge

Introduction

- **A large-scale, multi-label problem**
  - One protein can be associated with multiple GO terms By regrading
    - each GO term as a label
    - each protein as an instance
    - AFP can be deemed as a large-scale, **multi-label problem**
  - GO is a directed acyclic graph (DAG), meaning that for one protein annotated by one GO term, all ancestor GO terms in GO can be also assigned ? 여기 잘 이해가..?
  - In fact, one human protein is currently associated **with 47 GO terms on average**, according to Gene Ontology Annotation (GOA) Database (Dec 2020)

Introduction

- **Previous works**
  - GOLabeler
    - Sequence-based AFP method
    - Achieved the first place in CAFA3 on no-knowledge benchmark in terms of F_max in all three GO domains
    - Utilizes learning to rank (LTR) to integrate multiple types of sequence information
      - such as sequence homology, protein domain and family
      - to rank the candidate GO terms for a given protein (가장 순위가 높은 GO term 부여를 위해)
    - sequence information is insufficient to characterize protein functions

Introduction

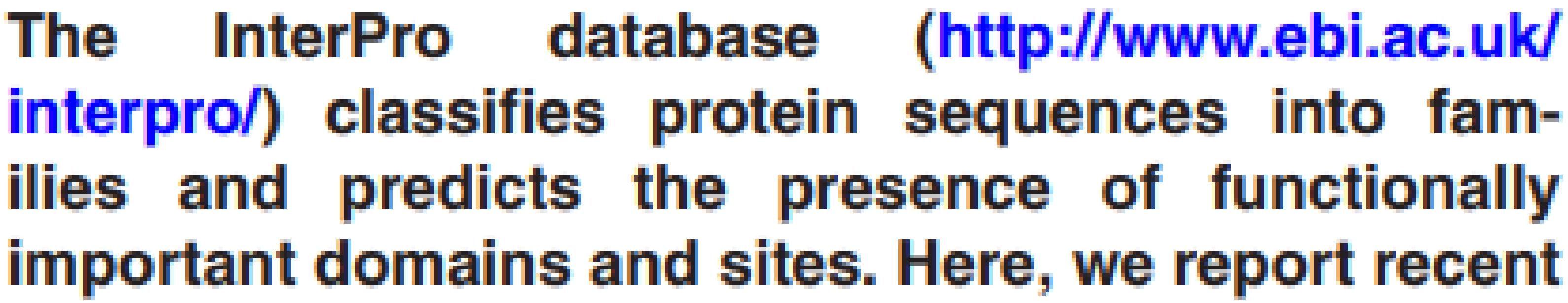
- **Previous works**
  - NetGO
    - Promising idea to improve AFP is that proteins connected in a protein network (e.g. protein-protein interaction or metabolic network) like **to share the same functions**
    - Keeping the LTR framework of GOLabeler, to improve the performance of GOLabeler by massive network information in STRING
    - NetGO, the component method on networks considers **only neighbors of a test protein** (i.e. low-order information) in given networks, meaning that high-order information in protein networks are ignored

Introduction

- **DeepGraphGO**
  - A semi-supervised, deep learning method, which takes the advantages of **both protein sequence** and **network information** through graph neural network (GNN)
  - has the following three notable features:
    1. **InterPro** for representation vector
       - The input of representation vectors
    2. **Multiple graph convolutional neural** (GCN) layers
       - Multiple GCN layers allow to capture high-order information among nodes (proteins)
    3. **Multispecies** strategy
       - Used proteins of all species for training only one single model

Introduction

- **InterPro**
  - The input of representation vectors (of nodes/proteins), trained by GNN, is generated from InterPro (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6323941/pdf/gky1100.pdf)
  - A protein domain and family database, combines 14 different databases



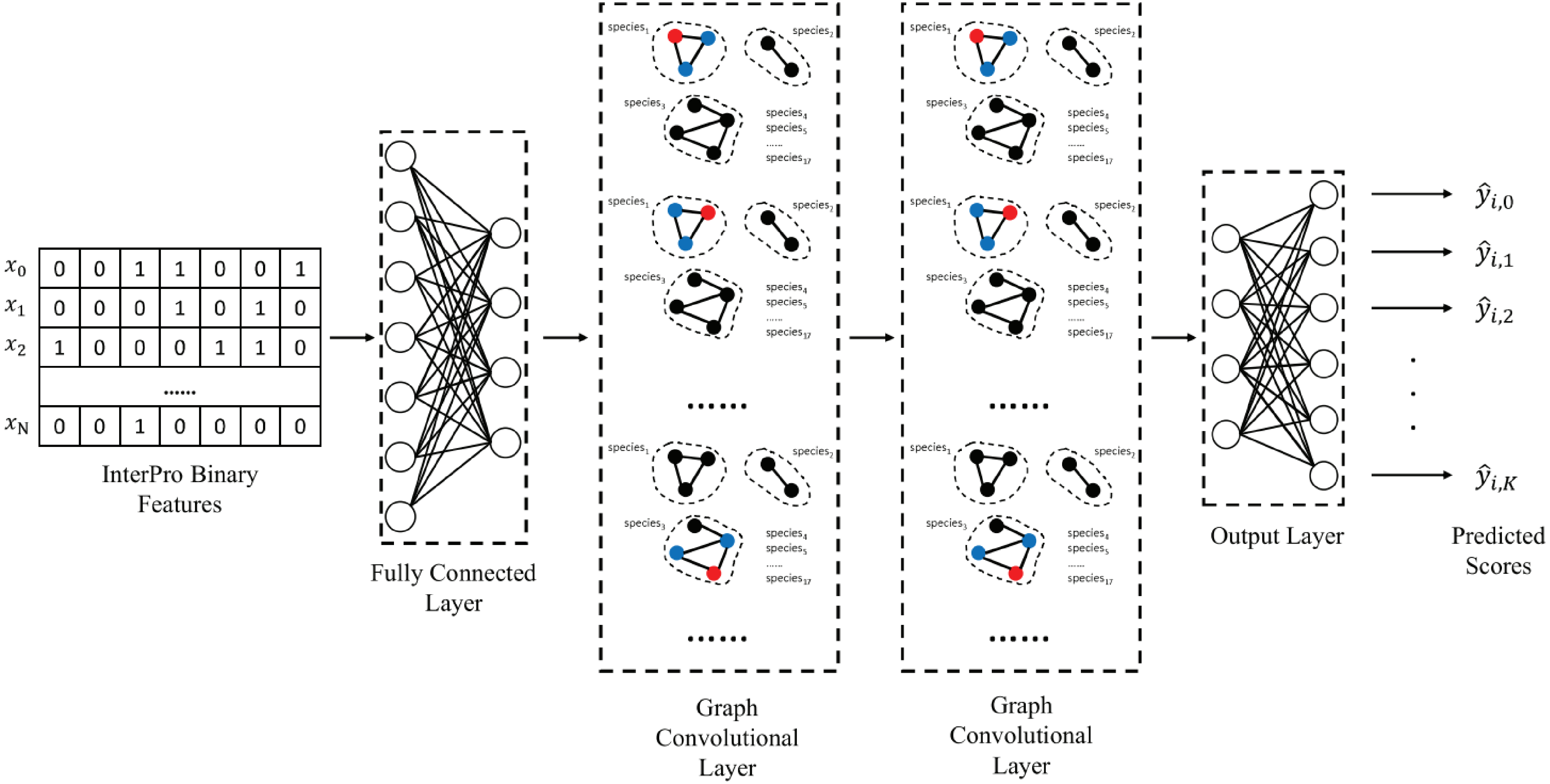
The InterPro database (http://www.ebi.ac.uk/interpro/) classifies protein sequences into families and predicts the presence of functionally important domains and sites. Here, we report recent

  - Provides many types of functional information, such as family, domain and motifs.
  - The features extracted from InterPro were successfully used in GOLabeler and NetGO as well

2. Materials and methods

Materials and methods

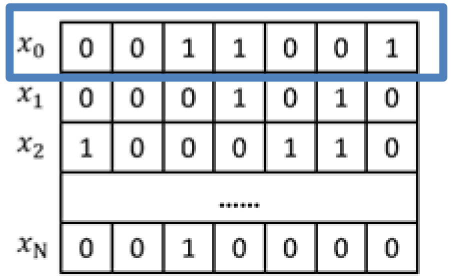
- **Overview**



x_0	0	0	1	1	0	0	1
x_1	0	0	0	1	0	1	0
x_2	1	0	0	0	1	1	0
	......						
x_N	0	0	1	0	0	0	0

Materials and methods

- **Overview**
  - two inputs:
    1. graph G (protein network) with N nodes (proteins) or weighted adjacency matrix $A^2 \in \mathbb{R}^{N \times N}$ (edge weights range between 0 and 1).
    2. (ii) N binary feature vectors, generated by InterProScan (Jones et al., 2014) for N proteins based on InterPro, where each element shows the **presence/absence** of a protein domain/family/motif



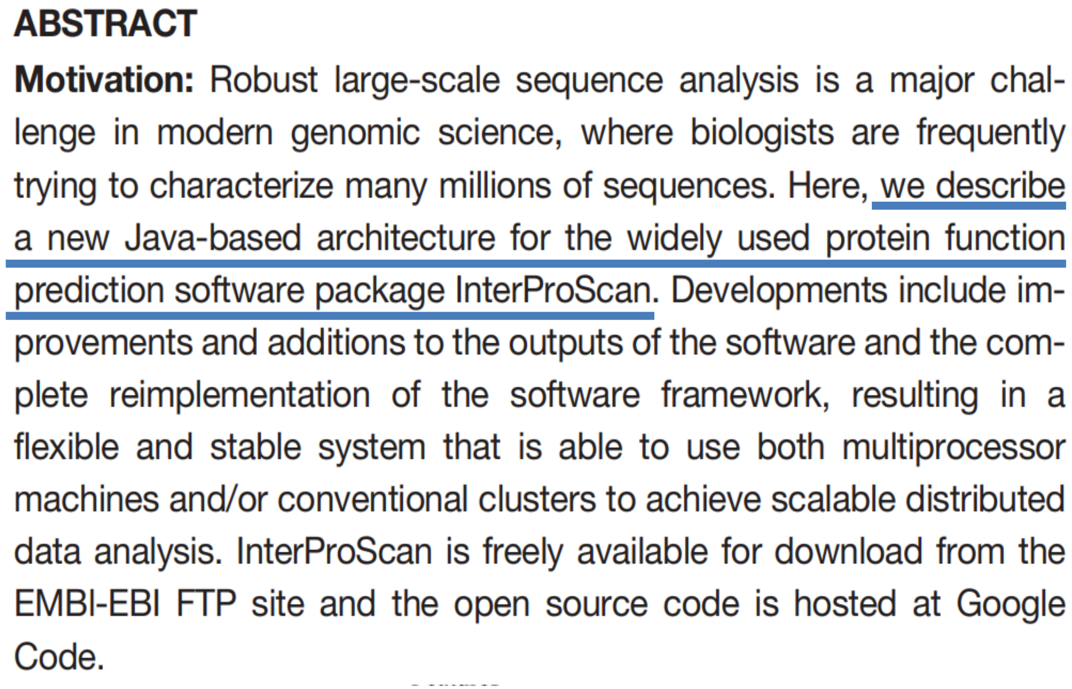
x_0	0	0	1	1	0	0	1
x_1	0	0	0	1	0	1	0
x_2	1	0	0	0	1	1	0
	......						
x_N	0	0	1	0	0	0	0

InterPro Binary Features

Protein 의 domain, family, motif 가 주어져있을 때 (구조라고 생각하면 편함) 특정 protein x_0 가 그 domain 에 해당한다면 1 아니라면 0

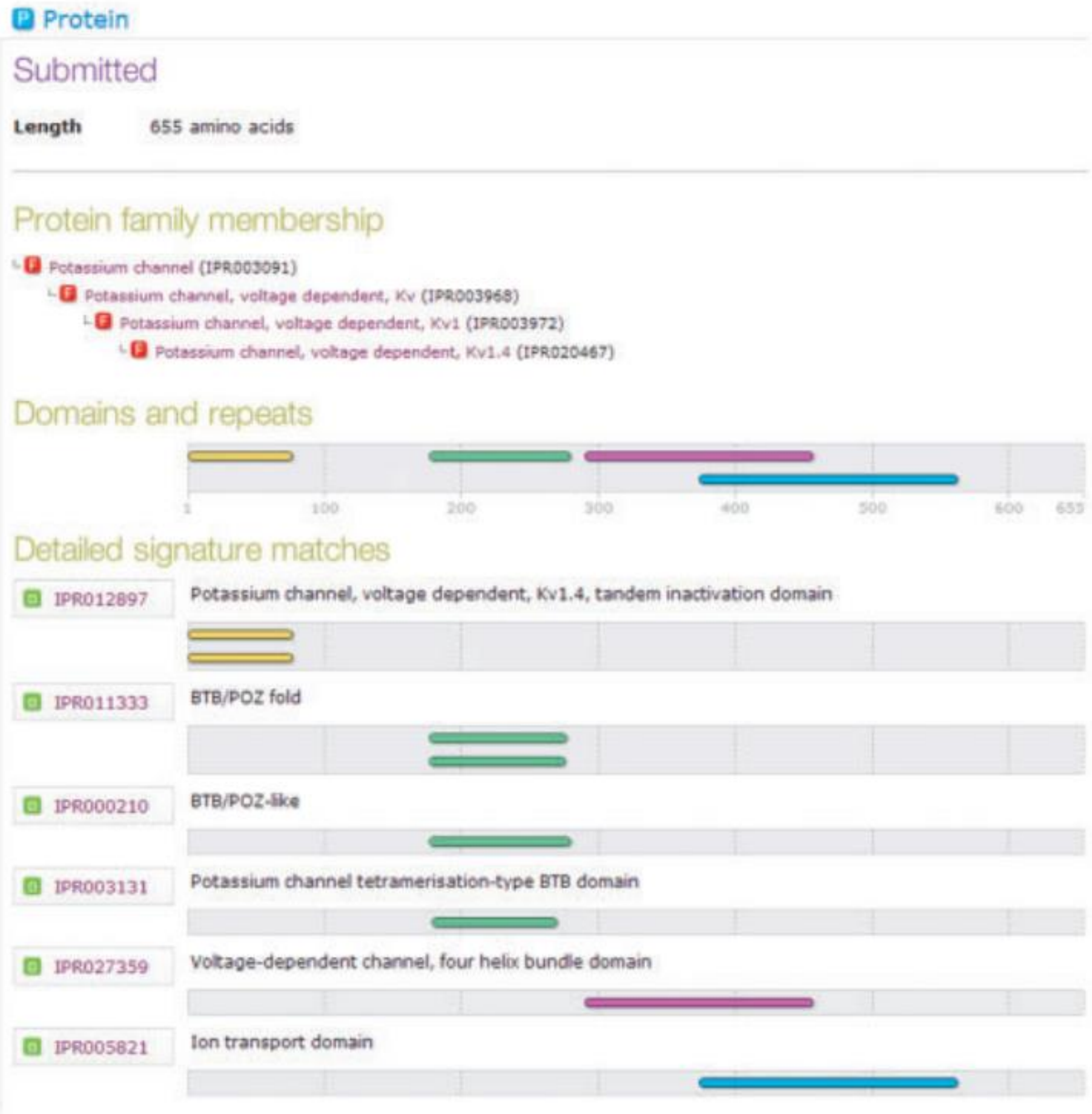
Materials and methods

- **Jones et al., 2014**
  - InterProScan 5: genome-scale protein function classification (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3998142/pdf/btu031.pdf)



ABSTRACT

Motivation: Robust large-scale sequence analysis is a major challenge in modern genomic science, where biologists are frequently trying to characterize many millions of sequences. Here, we describe a new Java-based architecture for the widely used protein function prediction software package InterProScan. Developments include improvements and additions to the outputs of the software and the complete reimplementation of the software framework, resulting in a flexible and stable system that is able to use both multiprocessor machines and/or conventional clusters to achieve scalable distributed data analysis. InterProScan is freely available for download from the EMBI-EBI FTP site and the open source code is hosted at Google Code.



P Protein

Submitted

Length 655 amino acids

Protein family membership

- F Potassium channel (IPR003091)
  - F Potassium channel, voltage dependent, Kv (IPR003968)
    - F Potassium channel, voltage dependent, Kv1 (IPR003972)
      - F Potassium channel, voltage dependent, Kv1.4 (IPR020467)

Domains and repeats

Detailed signature matches

IPR012897	Potassium channel, voltage dependent, Kv1.4, tandem inactivation domain
IPR011333	BTB/POZ fold
IPR000210	BTB/POZ-like
IPR003131	Potassium channel tetramerisation-type BTB domain
IPR027359	Voltage-dependent channel, four helix bundle domain
IPR005821	Ion transport domain

Materials and methods

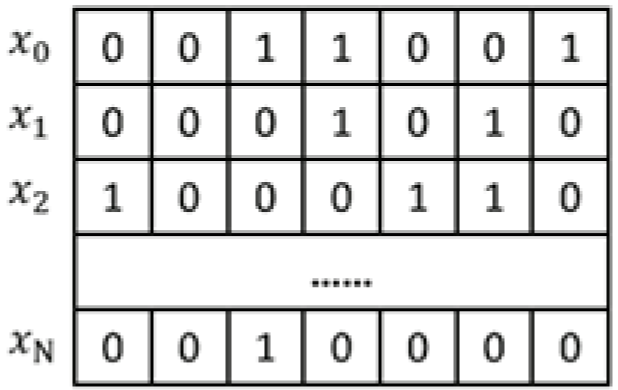
- **Procedure**

1. Input (fully connected) layer: **the binary feature vector of each protein** is transformed into a non-binary vector, to be used as the initial representation vector
2. Graph convolutional (GCN) layer: updates the representation vector of each node (protein) to capture **high-order information** through graph edges, by renewing the vector using those of neighboring nodes
3. Output (fully connected) layer: predicts scores of GO terms for each protein

Materials and methods

- **Input layer**

1. For protein p_i, InterProScan enerates a binary feature vector $x_i \in \{0,1\}^m$, where m is the number of signatures (domains and families in InterPro) related to at least one of N proteins in graph G ($m =$ 그래프 G에 있는 N개의 단백질 중 적어도 하나와 관련된 시그니처의 수), and the jth element of $x_i, x_{i,j}$, indicates if the jth signature belongs to p_i
2. 가로축 : 단백질 p_i, 세로축 : 특정 signature 가 단백질 p_i 에 있으면 1 없으면 0
3. 아마도 아래의 matrix 는 그래프 하나에 있는 단백질에 대한 것을 나타낸 듯?



x_0	0	0	1	1	0	0	1
x_1	0	0	0	1	0	1	0
x_2	1	0	0	0	1	1	0
	......						
x_N	0	0	1	0	0	0	0

InterPro Binary Features

3. Experiments

Experiments

- **Datasets**
  - Protein sequences:
    - Downloaded protein sequences from UniProt (https://www.uniprot.org/downloads)
  - Protein networks:
    - version 11.0 of STRING (https://string-db.org/) (Szklarczyk et al., 2019)
    - This database covers around 24.6 million proteins from 5090 organisms with more than two billion interactions in total
  - GO terms
    - Downloaded from SwissProt1 (Boutet et al., 2016), GOA (http://www.ebi.ac.uk/GOA) (Huntley et al., 2015) and GO (http://geneontology.org/page/download-annotations) (Ashburner et al., 2000)
    - Extracted all experimental annotations in: 'IDA', 'IPI', 'EXP', 'IGI', 'IMP', 'IEP', 'IC' or 'TA'. All are combined to generate an annotation dataset (https://wiki.geneontology.org/index.php/Inferred_from_Genetic_Interaction_(IGI))

Experiments

- **Datasets (Training, Validation, Test)**
  - Training : All data experimentally annotated before Jan. 2018.
  - Validation : All no-knowledge proteins experimentally annotated from January to December 2018.
  - Testing : All no-knowledge proteins experimentally annotated from January 2019 to January 2020.

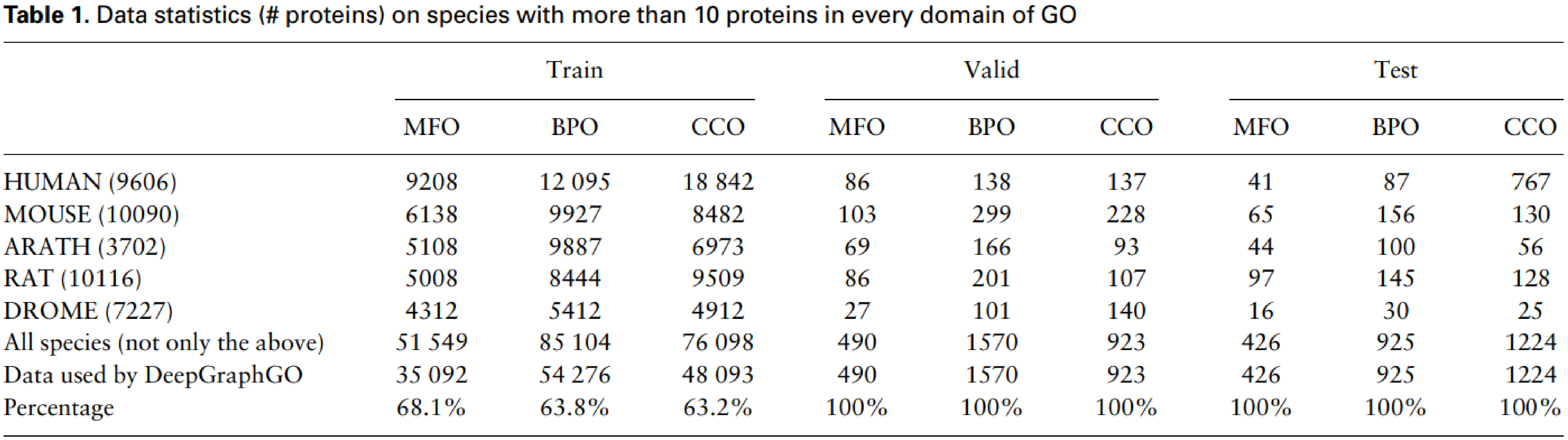


Table 1. Data statistics (# proteins) on species with more than 10 proteins in every domain of GO

	Train			Valid			Test		
	MFO	BPO	CCO	MFO	BPO	CCO	MFO	BPO	CCO
HUMAN (9606)	9208	12 095	18 842	86	138	137	41	87	767
MOUSE (10090)	6138	9927	8482	103	299	228	65	156	130
ARATH (3702)	5108	9887	6973	69	166	93	44	100	56
RAT (10116)	5008	8444	9509	86	201	107	97	145	128
DROME (7227)	4312	5412	4912	27	101	140	16	30	25
All species (not only the above)	51 549	85 104	76 098	490	1570	923	426	925	1224
Data used by DeepGraphGO	35 092	54 276	48 093	490	1570	923	426	925	1224
Percentage	68.1%	63.8%	63.2%	100%	100%	100%	100%	100%	100%

Experiments

- **Competing methods**
  - 실험 : 2개의 서로 다른 타입의 Evaluation method 를 사용
    - Protein-centric : 각각 protein 에 annotated 되어 있는 GO term 을 측정
    - GO term-centric : 각각 GO term에 해당하는 protein 을 측정
  - Protein-centric : DeepGOCNN, DeepGOPlus and three most important components of NetGO: BLAST-KNN, Net-KNN and LR-InterPro
  - GO term-centric : DeepNF, clusDCA and GeneMANIA

  - Trained DeepGraphGO for MFO, BPO and CCO separately
  - # of layers = 2 (이전에 더 깊은 layer 를 했으나 별 효과도 없고 computational cost 만 많이 잡아 먹었다고 나와있음)

Experiments

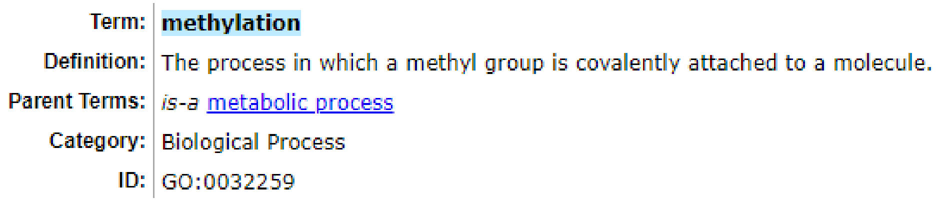
- **Performance evaluation metrics**
  - Three evaluation metrics F_{max}, AUPR, M-AUPR

1. F_{max} : Protein-centric (main evaluation metric)

2. AUPR : Widely used for performance evaluation of multi-label classification including AFP

3. M-AUPR : GO term-centric, being widely used by network-based methods

Experiments



Term:	**methylation**
Definition:	The process in which a methyl group is covalently attached to a molecule.
Parent Terms:	*is-a* metabolic process
Category:	Biological Process
ID:	GO:0032259

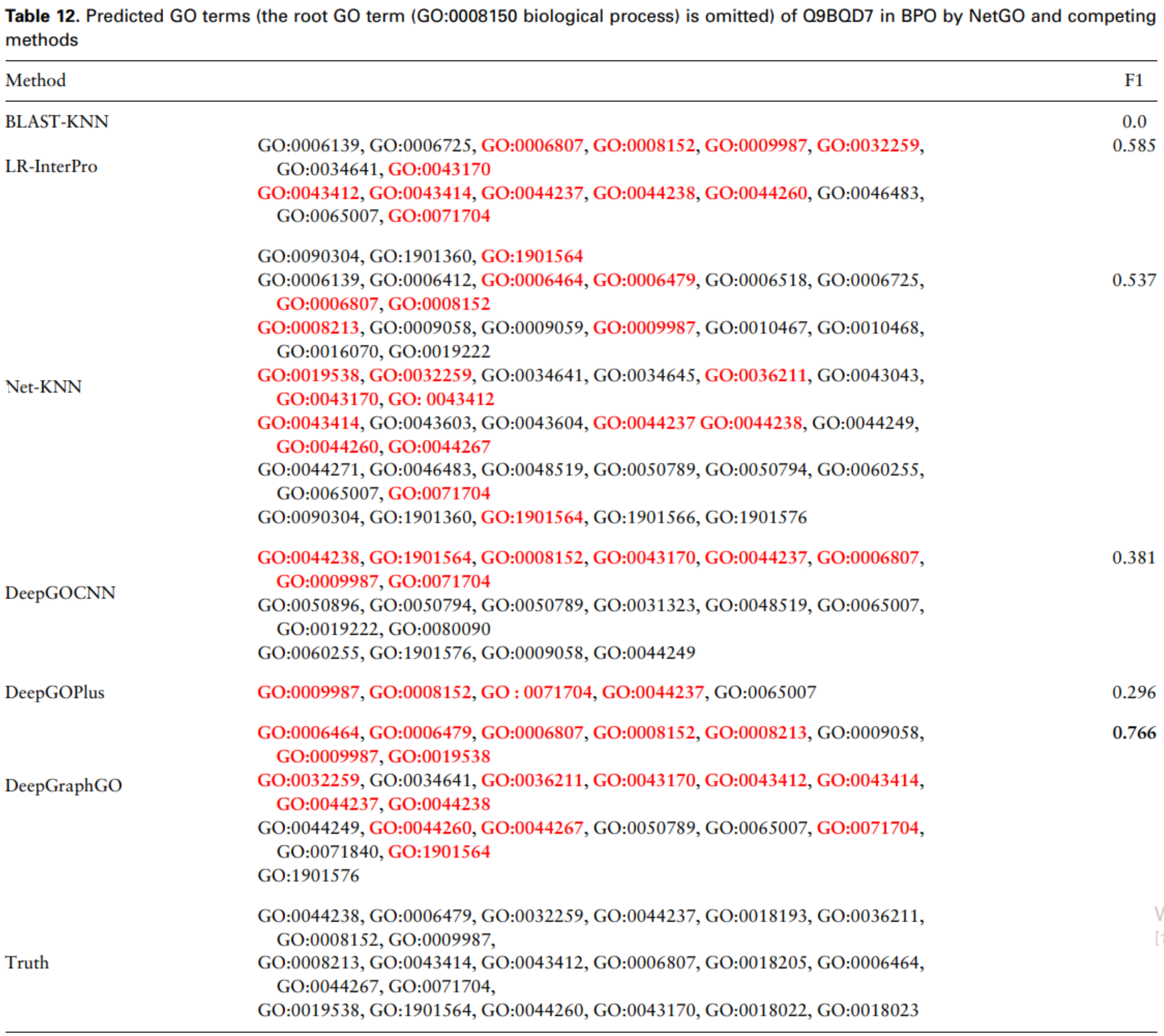


Table 12. Predicted GO terms (the root GO term (GO:0008150 biological process) is omitted) of Q9BQD7 in BPO by NetGO and competing methods

Method		F1
BLAST-KNN		0.0
LR-InterPro	GO:0006139, GO:0006725, GO:0006807, GO:0008152, GO:0009987, GO:0032259, GO:0034641, GO:0043170 GO:0043412, GO:0043414, GO:0044237, GO:0044238, GO:0044260, GO:0046483, GO:0065007, GO:0071704 GO:0090304, GO:1901360, GO:1901564	0.585
Net-KNN	GO:0006139, GO:0006412, GO:0006464, GO:0006479, GO:0006518, GO:0006725, GO:0006807, GO:0008152 GO:0008213, GO:0009058, GO:0009059, GO:0009987, GO:0010467, GO:0010468, GO:0016070, GO:0019222 GO:0019538, GO:0032259, GO:0034641, GO:0034645, GO:0036211, GO:0043043, GO:0043170, GO: 0043412 GO:0043414, GO:0043603, GO:0043604, GO:0044237 GO:0044238, GO:0044249, GO:0044260, GO:0044267 GO:0044271, GO:0046483, GO:0048519, GO:0050789, GO:0050794, GO:0060255, GO:0065007, GO:0071704 GO:0090304, GO:1901360, GO:1901564, GO:1901566, GO:1901576	0.537
DeepGOCNN	GO:0044238, GO:1901564, GO:0008152, GO:0043170, GO:0044237, GO:0006807, GO:0009987, GO:0071704 GO:0050896, GO:0050794, GO:0050789, GO:0031323, GO:0048519, GO:0065007, GO:0019222, GO:0080090 GO:0060255, GO:1901576, GO:0009058, GO:0044249	0.381
DeepGOPlus	GO:0009987, GO:0008152, GO : 0071704, GO:0044237, GO:0065007	0.296
DeepGraphGO	GO:0006464, GO:0006479, GO:0006807, GO:0008152, GO:0008213, GO:0009058, GO:0009987, GO:0019538 GO:0032259, GO:0034641, GO:0036211, GO:0043170, GO:0043412, GO:0043414, GO:0044237, GO:0044238 GO:0044249, GO:0044260, GO:0044267, GO:0050789, GO:0065007, GO:0071704, GO:0071840, GO:1901564 GO:1901576	**0.766**
Truth	GO:0044238, GO:0006479, GO:0032259, GO:0044237, GO:0018193, GO:0036211, GO:0008152, GO:0009987, GO:0008213, GO:0043414, GO:0043412, GO:0006807, GO:0018205, GO:0006464, GO:0044267, GO:0071704, GO:0019538, GO:1901564, GO:0044260, GO:0043170, GO:0018022, GO:0018023	

4. Conclusion

Conclusion

- **Conclusion**
  - Designed an end-to-end, graph neural network-based model, DeepGraphGO,
  - Challenging AFP problem, to make the most of both protein sequence and protein network information
  - DeepGraphGO uses ‘multi-species strategy’, which allows only one single model to be trained **by using proteins of all species**
  - Possible future work would be to build a single model for AFP, which can incorporate all kind of protein information including sequence, structure and network.

Q & A

Thank You!