Syracuse University

Dublin Core Data Institute
An International Conference
19th October, 2024 Toronto, Canada

Using Gene Ontology in ML Models: The Design and Creation of Bio-Knowledgebases

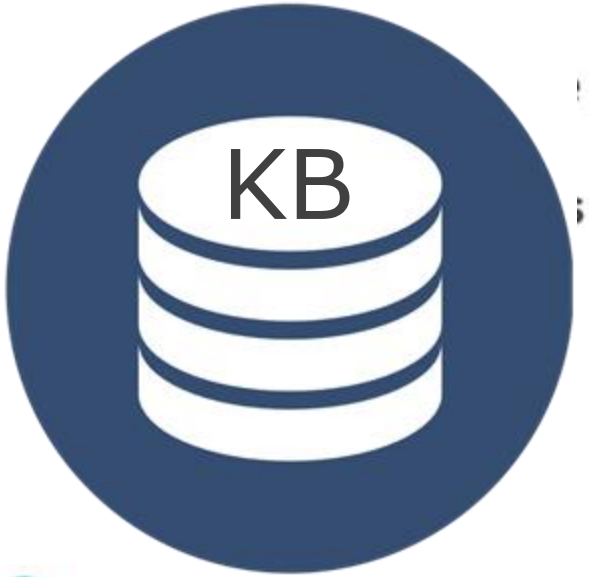
Qiaoyi (Joy) Liu, PhD Student

Jian Qin, Professor

Syracuse University

Knowledgebases as a kind of modern knowledge organization systems for supporting knowledge networks using computational power

Knowledge organization



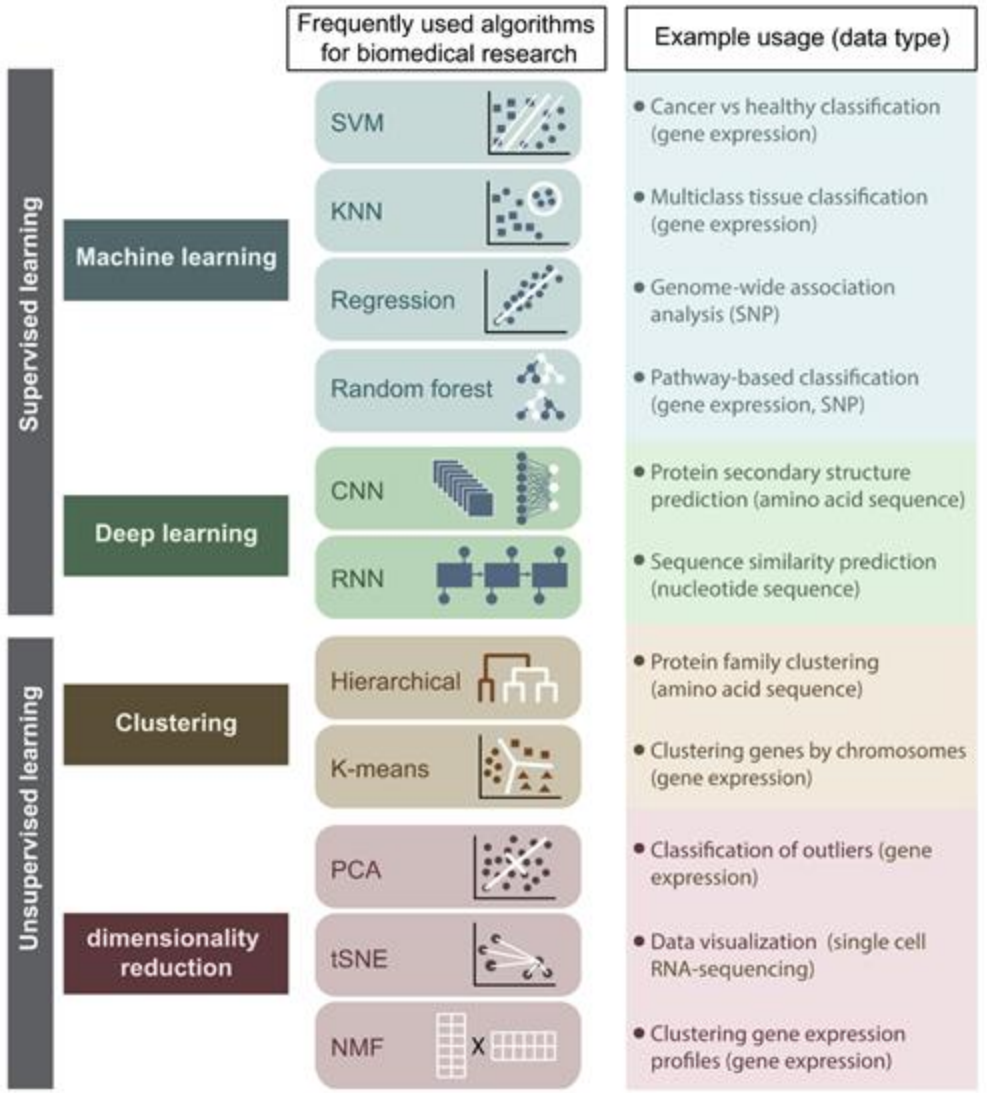
Knowledge annotation

Biomedical Knowledgebases

- Structural biomedical domain knowledge fusion.
- Flexibly organize semantic information e.g. chemical compounds, genes, proteins, drugs, diseases to integrate with external resources.
- Widely used for disease identification and clinical diagnosis.

Machine Learning in Biomedical & Bioinformatics Research

- New era of biology
- Representation, storage, management, analysis and investigation of various data types
- Sophisticated algorithms and computational tools

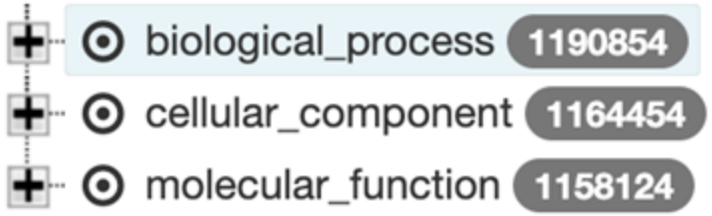


Auslander et al., 2021.

Gene Ontology (GO)

Structured, computer-accessible representation of :

- Genes and gene products
- Ontology: GO terms, synonym, definition
- Annotations



https://amigo.geneontology.org/amigo/dd_browse

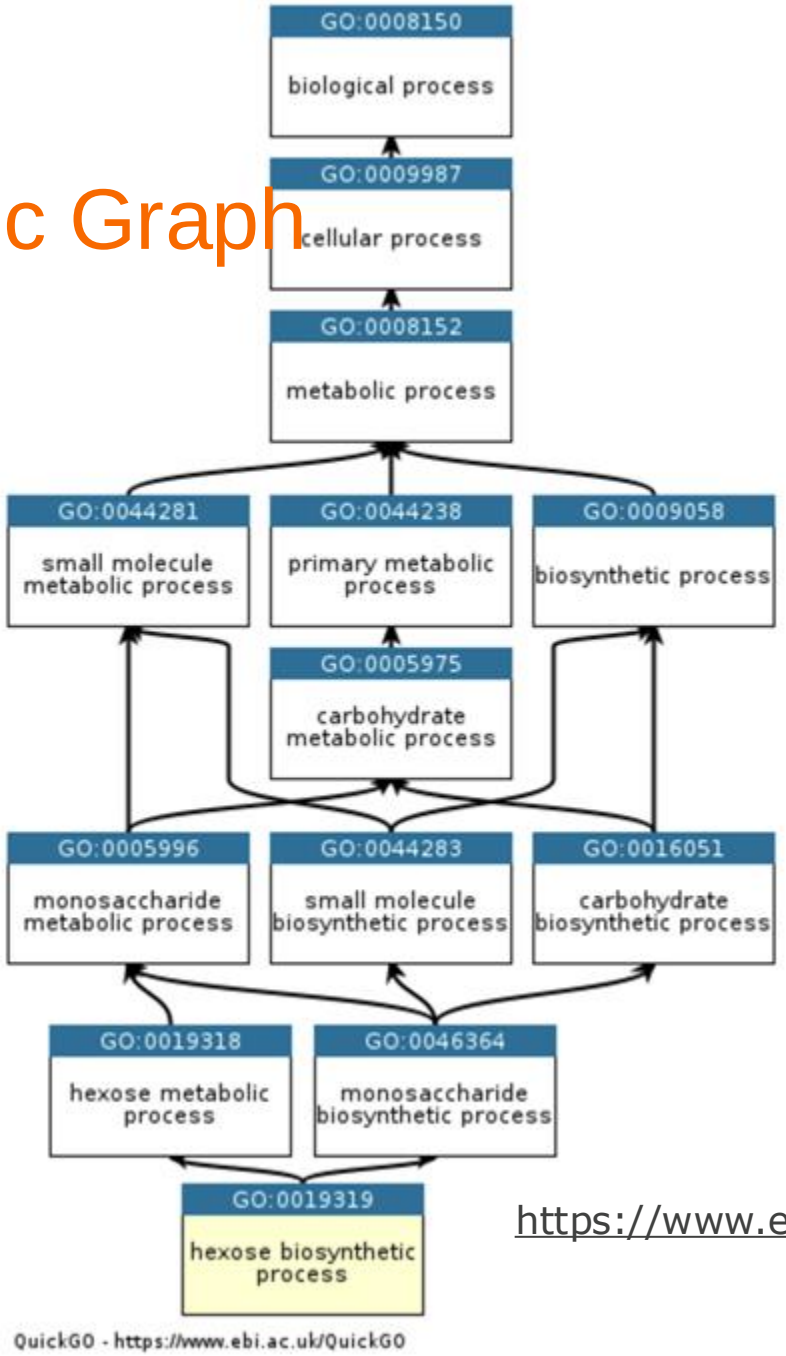
hexose biosynthetic process

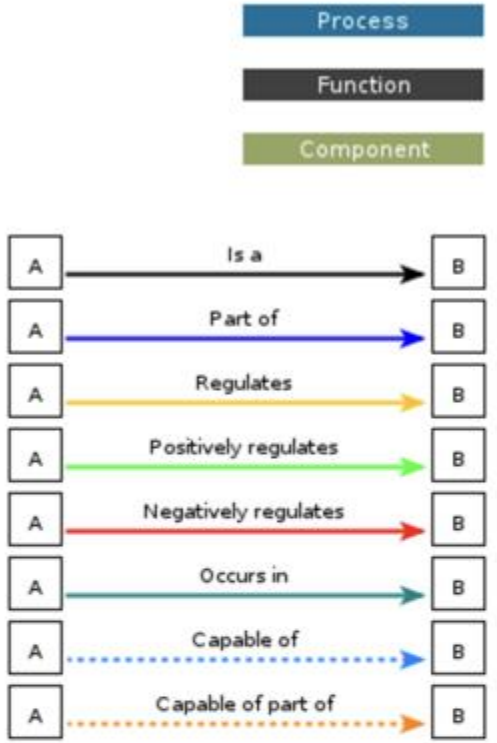
Term Information

Feedback

Accession	GO:0019319
Name	hexose biosynthetic process
Ontology	biological_process
Synonyms	hexose anabolism, hexose biosynthesis, hexose formation, hexose synthesis
Alternate IDs	None
Definition	The chemical reactions and pathways resulting in the formation of hexose, any monosaccharide with a chain of six carbon atoms in the molecule. *Source:* ISBN:0198506732
Comment	None
History	See term history for GO:0019319 at QuickGO
Chem. react.	None
Subset	None
Related	Link to all **genes and gene products** annotated to hexose biosynthetic process (**excluding "regulates"**). Link to all direct and indirect **annotations** to hexose biosynthetic process (**excluding "regulates"**). Link to all direct and indirect **annotations download** (limited to first 10,000) for hexose biosynthetic process (**excluding "regulates"**).

GO Structure: Directed Acyclic Graph (DAG)

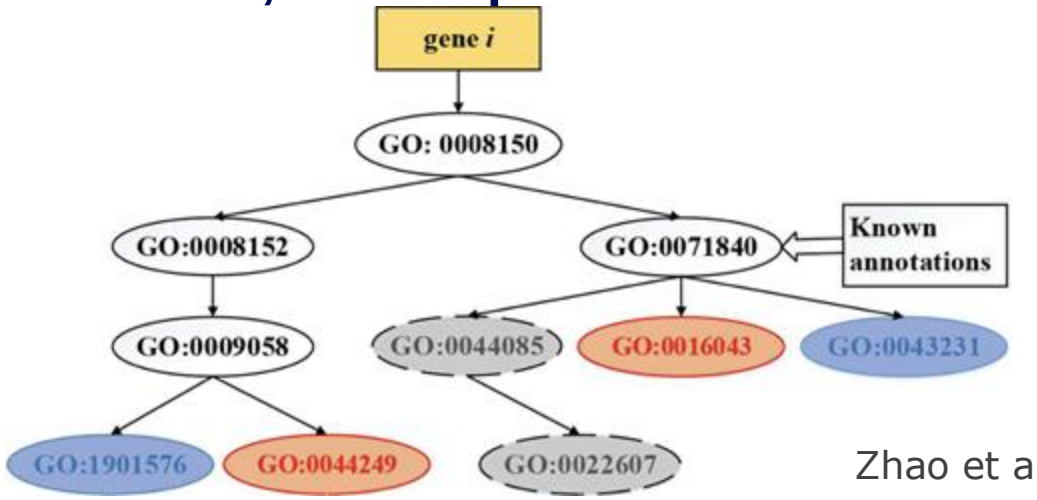




https://www.ebi.ac.uk/QuickGO/GTerm?id=GO:0019319

GO Annotation Data

- Associations between GO terms and genes or gene products
- Each annotation only represents one characteristic of the gene product
- One gene product, multiple GO annotations



Zhao et al., 2020.

Evidence Codes

- Experimental evidence
- Phylogenetic evidence
- Computational evidence
- Author statements
- Curatorial statements
- Automatically generated annotations

Experimental Evidence Codes		**Computational Analysis Evidence Codes**	
EXP	Inferred from Experiment	ISS	Inferred from Sequence or Structural Similarity
IDA	Inferred from Direct Assay	ISO	Inferred from Sequence Orthology
IPI	Inferred from Physical Interaction	ISA	Inferred from Sequence Alignment
IMP	Inferred from Mutant Phenotype	ISM	Inferred from Sequence Model
IGI	Inferred from Genetic Interaction	IGC	Inferred from Genomic Context
IEP	Inferred from Expression Pattern	RCA	Inferred from Reviewed Computational Analysis
Author Statement Evidence Codes		**Curator Statement Evidence Codes**	
TAS	Traceable Author Statement	IC	Inferred by Curator
NAS	Non-traceable Author Statement	ND	No biological Data available
Automatically-assigned Evidence Codes		**Obsolete Evidence Codes**	
IEA	Inferred from Electronic Annotation	NR	Not Recorded

Demo: Using BaseSet package to retrieve GO data

	elements	sets	DB	DB_Object_ID	Evidence_Code	With_From	DB_Object_Name	DB_Object_Type
1	URS0000001346_9606	GO:0006412	RNAcentral	URS0000001346_9606	IEA	GO:0030533	Homo sapiens (human) tRNA-Lys	tRNA
2	URS0000001346_9606	GO:0030533	RNAcentral	URS0000001346_9606	IEA	Rfam:RF00005	Homo sapiens (human) tRNA-Lys	tRNA
3	URS000000192A_9606	GO:0016442	RNAcentral	URS000000192A_9606	IEA	Rfam:RF00951	Homo sapiens (human) MIR1302-2 host gene (MIR13...	lnc_RNA
4	URS000000192A_9606	GO:0035195	RNAcentral	URS000000192A_9606	IEA	Rfam:RF00951	Homo sapiens (human) MIR1302-2 host gene (MIR13...	lnc_RNA
5	URS00000019BC_9606	GO:0000244	RNAcentral	URS00000019BC_9606	IEA	Rfam:RF00026	Homo sapiens (human) snRNA-U6-related	snRNA
6	URS00000019BC_9606	GO:0000353	RNAcentral	URS00000019BC_9606	IEA	Rfam:RF00026	Homo sapiens (human) snRNA-U6-related	snRNA
7	URS00000019BC_9606	GO:0005688	RNAcentral	URS00000019BC_9606	IEA	Rfam:RF00026	Homo sapiens (human) snRNA-U6-related	snRNA
8	URS00000019BC_9606	GO:0030621	RNAcentral	URS00000019BC_9606	IEA	Rfam:RF00026	Homo sapiens (human) snRNA-U6-related	snRNA
9	URS00000019BC_9606	GO:0046540	RNAcentral	URS00000019BC_9606	IEA	Rfam:RF00026	Homo sapiens (human) snRNA-U6-related	snRNA
10	URS0000001A7A_9606	GO:0016442	RNAcentral	URS0000001A7A_9606	IEA	Rfam:RF00027|Rfam:RF00027	Homo sapiens (human) microRNA hsa-mir-625 precu...	primary_transcrip
11	URS0000001A7A_9606	GO:0035195	RNAcentral	URS0000001A7A_9606	IEA	Rfam:RF00027|Rfam:RF00027	Homo sapiens (human) microRNA hsa-mir-625 precu...	primary_transcrip
12	URS0000003515_9606	GO:0003735	RNAcentral	URS0000003515_9606	IEA	Rfam:RF01959	Homo sapiens (human) 12S ribosomal RNA	rRNA
13	URS0000003515_9606	GO:0005840	RNAcentral	URS0000003515_9606	IEA	Rfam:RF01959	Homo sapiens (human) 12S ribosomal RNA	rRNA
14	URS00000035FF_9606	GO:0003735	RNAcentral	URS00000035FF_9606	IEA	Rfam:RF00177	Homo sapiens (human) 12S ribosomal RNA	rRNA
15	URS00000035FF_9606	GO:0005840	RNAcentral	URS00000035FF_9606	IEA	Rfam:RF00177	Homo sapiens (human) 12S ribosomal RNA	rRNA

- Gene MatriX (GMX),
- GO Annotation File(GAF), or
- Open Biological and Biomedical Ontology Foundry (OBO)

Issues to consider with using GO data with ML:

- Incomplete and imbalanced GO annotations
- Under-established evaluation metrics for data quality
- Transformation of conceptual framework to reliable empirical data sources for Deep Learning and LLM
- Calculations of functional similarity between GO terms
- Compatibility with external knowledgebases and models

Reference

- Auslander, N., Gussow, A., & Koonin, E. (2021). Incorporating Machine Learning Into Bioinformatics Frameworks.
- Qin, J., & Liu, Q. (2024, January). Organizing Knowledge in Knowledgebases: A Case Study. In Knowledge Organization for Resilience in Times of Crisis: Challenges and Opportunities(pp. 393-400). Ergon-Verlag.
- Revilla Sancho L (2024). BaseSet: Working with Sets the Tidy Way. R package version 0.9.0.9002, https://docs.ropensci.org/BaseSet/, https://github.com/ropensci/BaseSet.
- The Gene Consortium, The Gene Ontology knowledgebase in 2023, Gene, 224(1) (2023) 1–14. https://doi.org/10.1093/genetics/iyad031.
- Zhao, Y., Wang, J., Chen, J., Zhang, X., Guo, M., & Yu, G. (2020). A literature review of gene function prediction by modeling gene ontology. Frontiers in genetics, 11, 400.

Syracuse University

Thank you!

Qiaoyi Liu
qliu11@syr.edu
Dr. Jian Qin
jqin@syr.edu