The coevolution of ontologies and knowledge-based analytics in bioinformatics

Robert Hoehndorf

King Abdullah University of Science and Technology

Co-evolution

Source: Wikipedia

Timeline

Biology/Bioinformatics

- First eukaryotes sequenced

Ontologies

- Gene Ontology

The origin of bio-ontologies, and GO (1999)

"Functional conservation requires a common language for annotation"

"The first comparison between two complete eukaryotic genomes (budding yeast and worm) revealed that a surprisingly large fraction of the genes in these two organisms displayed evidence of orthology"

"This astonishingly high degree of sequence and functional conservation presents both opportunities and challenges"

Ashburner et al., 2000

Solving the integration problem

Design decisions:

- taxonomy "to allow automatic transfers of annotation" between model organisms
- "to be able to organize, describe, query and visualize biological knowledge at vastly different stages of completeness"

Ashburner et al., 2000

Timeline

Biology/Bioinformatics

- First eukaryotes sequenced
- Microarray experiments

Ontologies

- Gene Ontology
- (Bio-)Ontological foundations

Enrichment analysis (2000–)

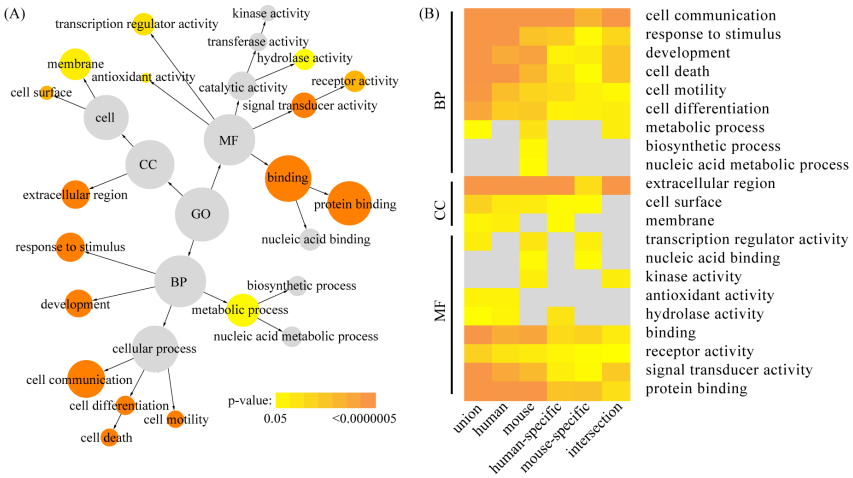
Characterizing gene sets

What characterizes a list of genes?

Enrichment analysis (2000–)

Characterizing gene sets

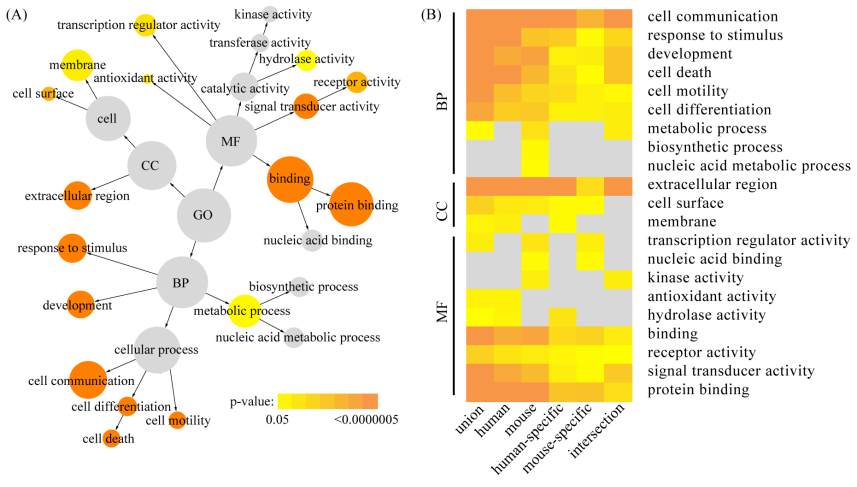
What characterizes a list of genes?

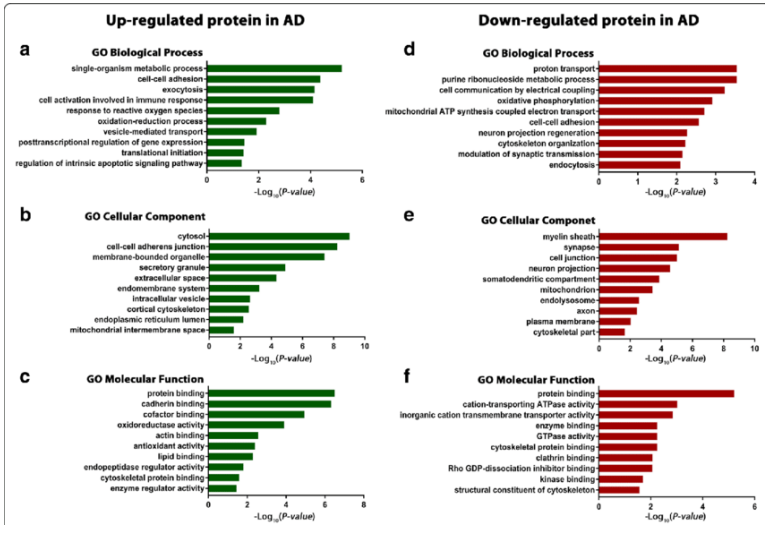


Enrichment analysis (2000–)

Characterizing gene sets

What characterizes a list of genes?

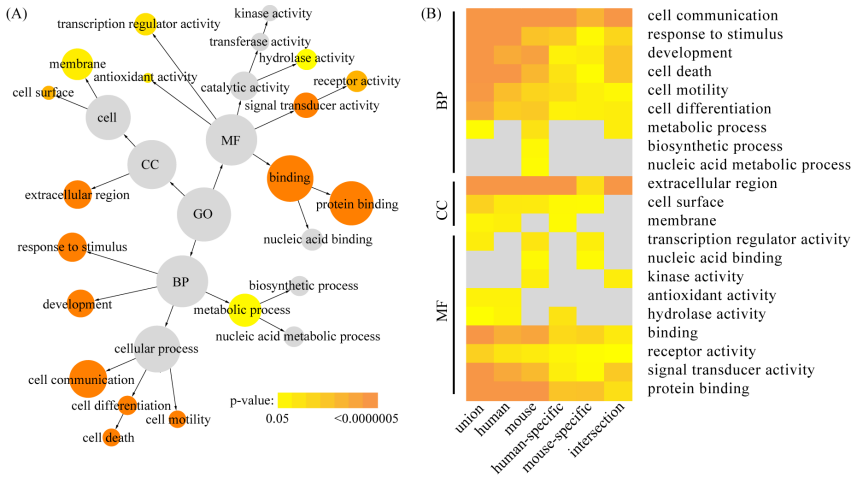


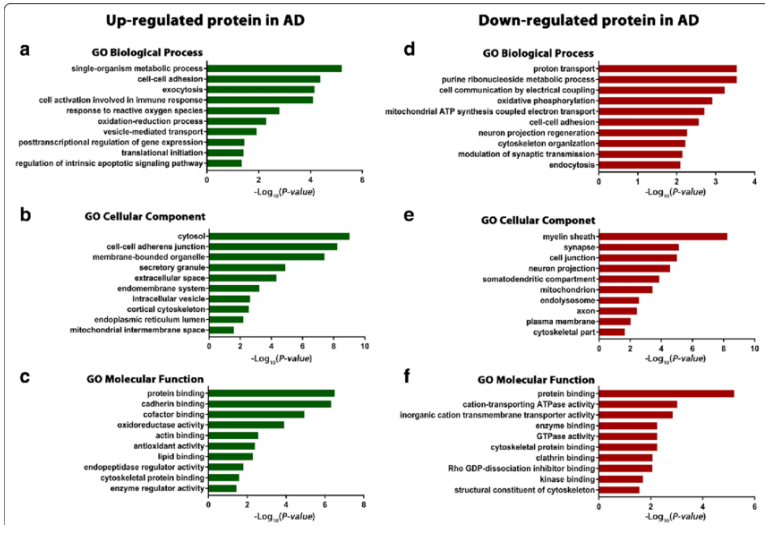


Enrichment analysis (2000–)

Characterizing gene sets

What characterizes a list of genes?

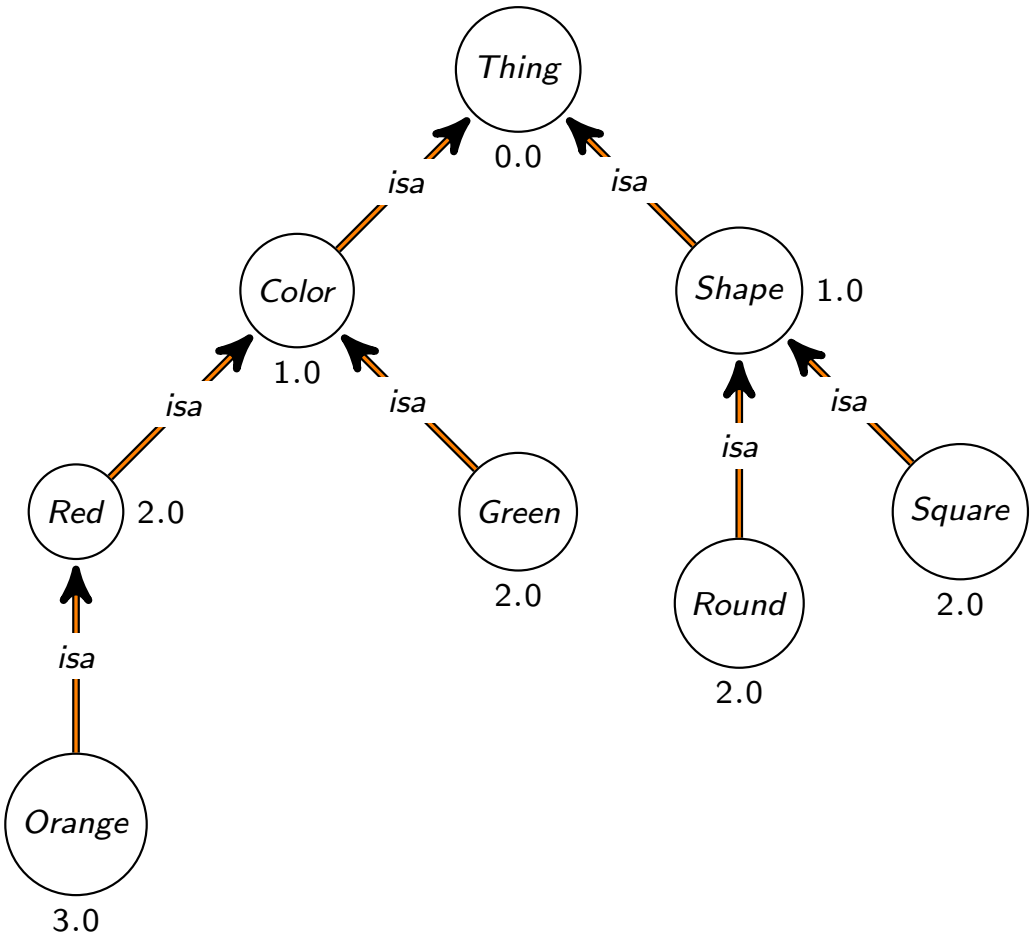




Successful interpretation relies on *accurate* propagation of annotations.

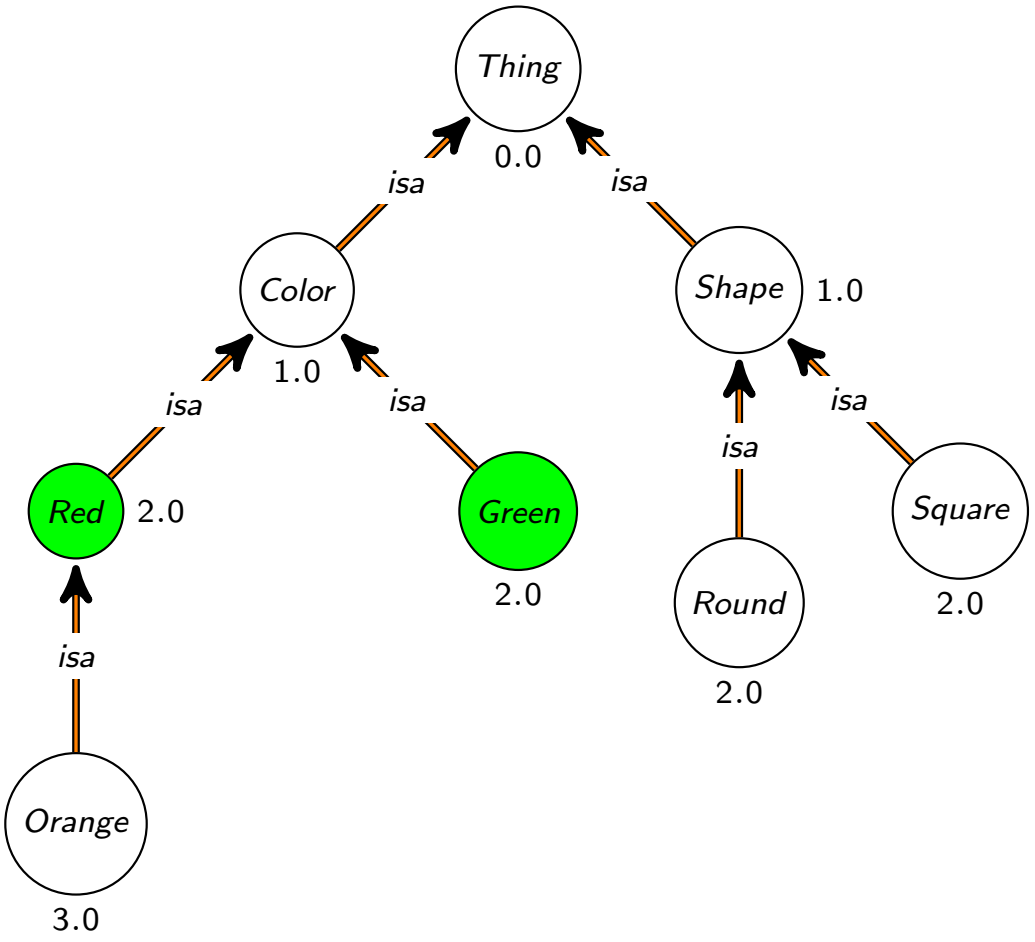
Semantic similarity (2003–)

Comparing proteins and sets of proteins



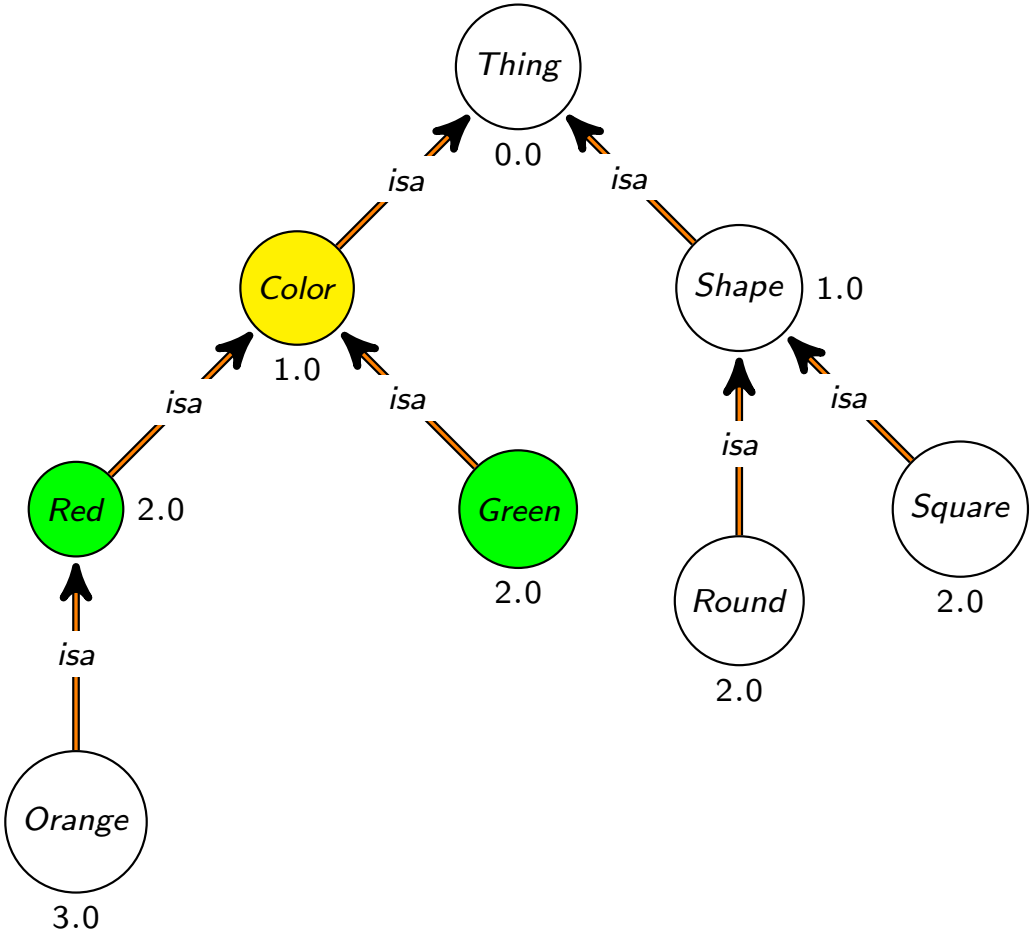
- Resnik 1995: similarity between x and y is the information content of the *most informative common ancestor*

Semantic similarity (2003–)



- Resnik 1995: similarity between x and y is the information content of the *most informative common ancestor*

Semantic similarity (2003–)



- Resnik 1995: similarity between x and y is the information content of the *most informative common ancestor*

Semantic similarity (2003–)

Semantic similarity

- retrieval on databases (Lord et al., 2003)
- predict disease genes ("guilt by association")
- differential diagnosis

All rely on accurate inferences in the ontology!

Fixing ontology problems (2003–2007)

- ontology-based analysis methods rely on *accurate* and *complete* inferences
  - the "true path rule"
  - aggregation of annotations along taxonomy/partonomy
- incorrect inferences result in incorrect scientific results/interpretations:
- is-a vs. part-of; necessary part vs contingent part; temporal dependency of part-of; causation vs part-of; absence, lacking parts

Timeline

Biology/Bioinformatics

- First eukaryotes sequenced
- Microarray experiments
- High throughput experiments

Ontologies

- Gene Ontology
- (Bio-)Ontological foundations
- Dealing with Big Data

Scaling up (2008–)

- high throughput technologies
- more data, more domains $\Rightarrow$ more ontologies
- manual curation no longer scales, too costly
- options:
  - automated construction of ontologies
  - ontology design patterns
  - lexical patterns
- relies on tools to *validate* constructed knowledge
  - automated reasoners

Scaling up (2008–)

Increasing the scale of bio-ontologies:

- modularization
  - MIREOT and associated tools
- light-weight reasoners
  - OWL 2 EL (Elk, Konklude)
  - only consider some (relatively weak) axioms

Consistency is no longer an attainable goal

OBO Foundry:

Ontology	Unsatisfiable Class Count
CHEBI	37
GO	565
OBI	34

Other:

Ontology Name	Unsatisfiable Class Count
Unified Phenotype Ontology (UPHENO)	106,126
Monarch Disease Ontology (MONDO)	97,619
Ontology for MIRNA Target (OMIT)	63,015
Molecular Process Ontology (MOP)	57,355
Name Reaction Ontology (RXNO)	57,330
Human Phenotype Ontology (HP)	46,075
Mammalian Phenotype Ontology (MP)	43,806
Cell Ontology (CL)	34,685
Ontology of Biological Attributes (OBA)	26,523
Ontology of Adverse Events (OAE)	20,566

Slater et al, 2020

Timeline

Biology/Bioinformatics

- First eukaryotes sequenced
- Microarray experiments
- High throughput experiments
- Clinical integration

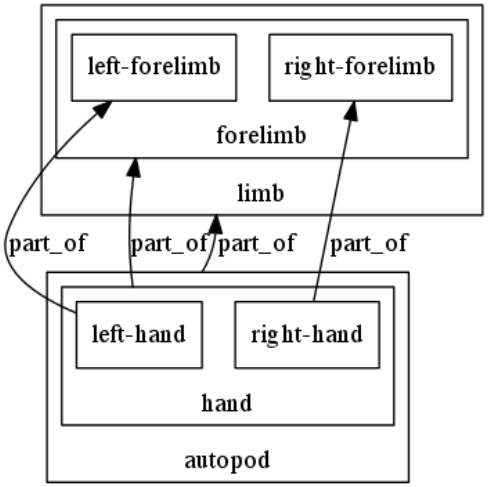
Ontologies

- Gene Ontology
- (Bio-)Ontological foundations
- Dealing with Big Data
- Knowledge graphs

Knowledge graphs (2015–)

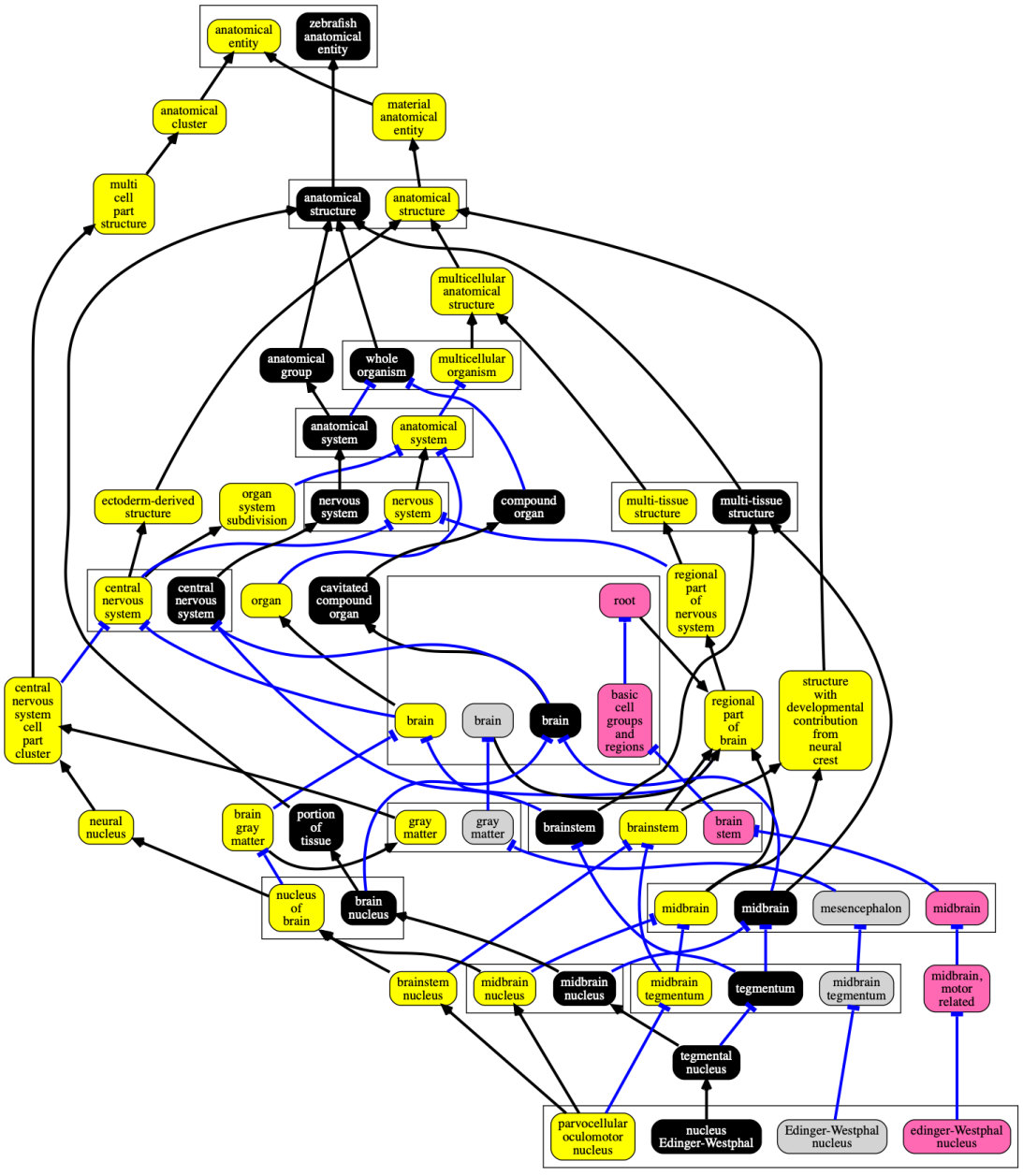
- Focus on linking, not semantics
- Ontologies are "projected" onto a graph
  - full circle

Axiom of Condition 1	Axiom or Triple(s) of Condition 2	Projected Triple(s)
$A \sqsubseteq \Box r.D$ or $\Box r.D \sqsubseteq A$	$D \equiv B \mid B_1 \sqcup \ldots \sqcup B_n \mid B_1 \sqcap \ldots \sqcap B_n$	$\langle A, r, B \rangle$ or $\langle A, r, B_i \rangle$ for $i \in 1, \ldots, n$
$\exists r.\top \sqsubseteq A$ (domain)	$\top \sqsubseteq \forall r.B$ (range)	
$A \sqsubseteq \exists r.\{b\}$	$B(b)$	
$r \sqsubseteq r'$	$\langle A, r', B \rangle$ has been projected	
$r' \equiv r^-$	$\langle B, r', A \rangle$ has been projected	
$s_1 \circ \ldots \circ s_n \sqsubseteq r$	$\langle A, s_1, C_1 \rangle \ldots \langle C_n, s_n, B \rangle$ have been projected	
$B \sqsubseteq A$	–	$\langle B, rdfs{:}subClassOf, A \rangle$ $\langle A, rdfs{:}subClassOf^-, B \rangle$
$A(a)$	–	$\langle a, rdf{:}type, A \rangle$ $\langle A, rdf{:}type^-, a \rangle$
$r(a, b)$	–	$\langle a, r, b \rangle$



Chen et al., 2021 (left). OBOGraphs Github (right)

Ontologies and knowledge graphs



Timeline

Biology/Bioinformatics

- First eukaryotes sequenced
- Microarray experiments
- High throughput experiments
- Clinical integration
- Machine learning and AI

Ontologies

- Gene Ontology
- (Bio-)Ontological foundations
- Dealing with Big Data
- Knowledge graphs
- ???

Timeline

Biology/Bioinformatics

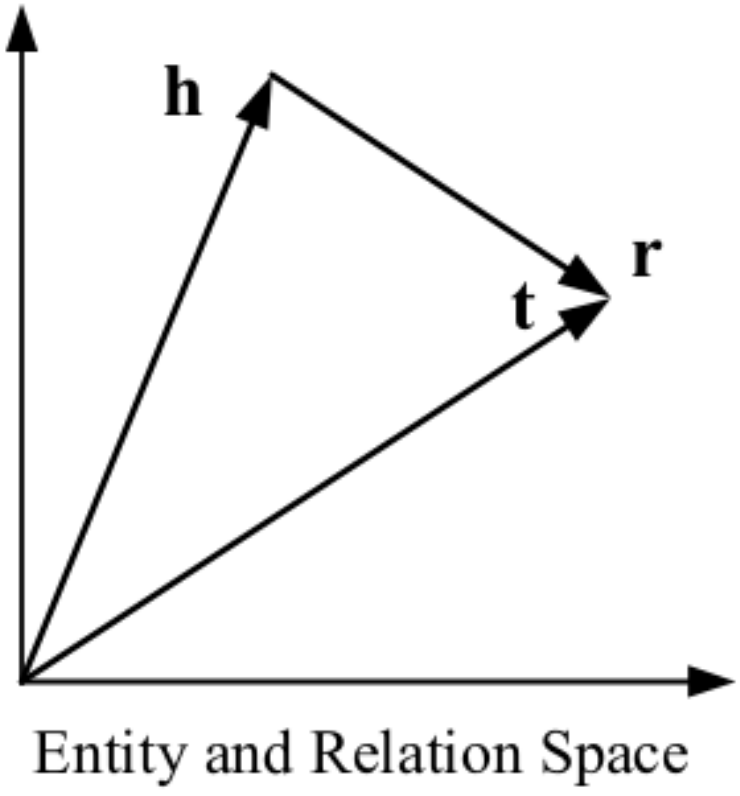
- First eukaryotes sequenced
- Microarray experiments
- High throughput experiments
- Clinical integration
- Machine learning and AI

Ontologies

- Gene Ontology
- (Bio-)Ontological foundations
- Dealing with Big Data
- Knowledge graphs
- ???

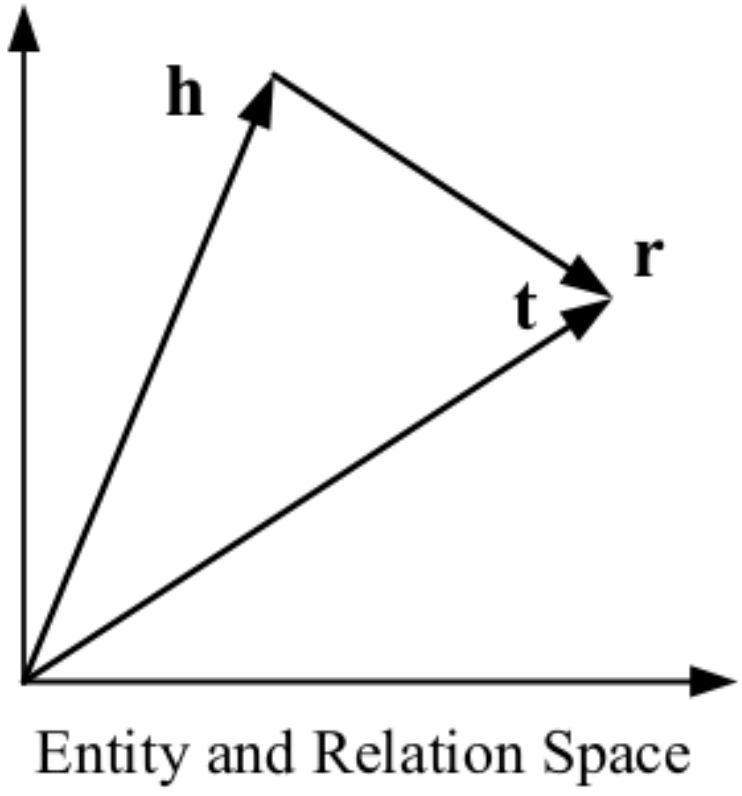
Will ontologies be only data providers for ML in biology?

Cool things to do with knowledge graphs



- head + rel = tail
- $\Rightarrow$ head + rel − tail = 0

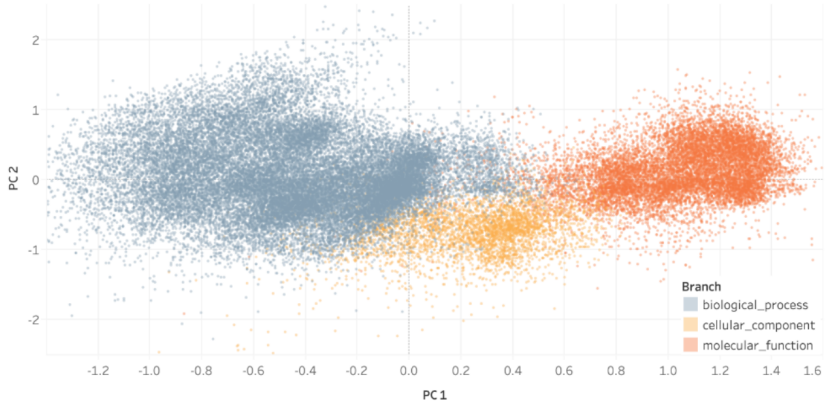
Cool things to do with knowledge graphs



- head + rel = tail
- $\Rightarrow$ head + rel − tail = 0
- head, rel, tail $\in \Re^n$
- for all triples in a graph

Cool things to do with knowledge graphs

Figure from: J Chen et al., 2021.

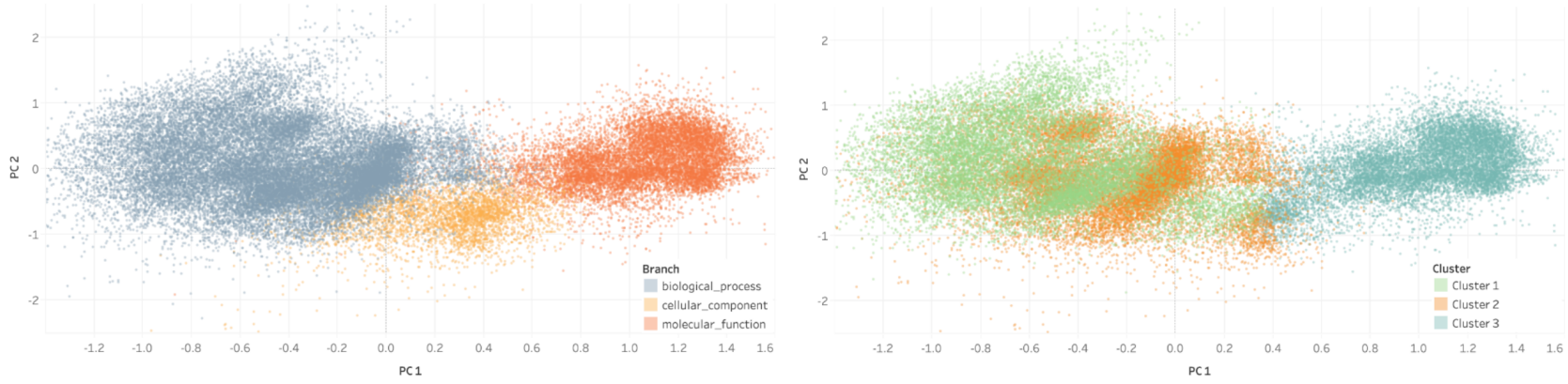


(a) Ontology branches.

(b) Three k-means clusters.

Cool things to do with knowledge graphs

Figure from: J Chen et al., 2021.



(a) Ontology branches. (b) Three k-means clusters.

Object property	Source type	Target type	Without reasoning		With reasoning	
			F-measure	AUC	F-measure	AUC
has target	Drug	Gene/Protein	0.94	0.97	0.94	0.98
has disease annotation	Gene/Protein	Disease	0.89	0.95	0.89	0.95
has side-effect*	Drug	Phenotype	0.86	0.93	0.87	0.94
has interaction	Gene/Protein	Gene/Protein	0.82	0.88	0.82	0.88
has function*	Gene/Protein	Function	0.85	0.95	0.83	0.91
has gene phenotype*	Gene/Protein	Phenotype	0.84	0.91	0.82	0.90
has indication	Drug	Disease	0.72	0.79	0.76	0.83
has disease phenotype*	Disease	Phenotype	0.72	0.78	0.70	0.77

Ontologies are more than knowledge graphs

Ontologies enable

- deductive inference
- complex, logical assertions and queries
- test of consistency
- model theory

Ontologies are more than knowledge graphs

Ontologies enable

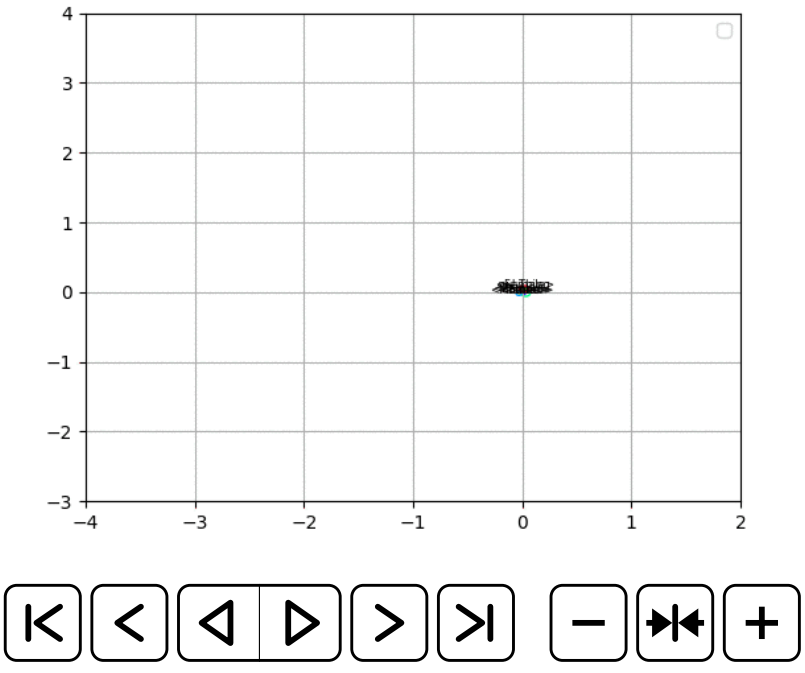
- deductive inference
- complex, logical assertions and queries
- test of consistency
- model theory

But there are not many methods in machine learning that can utilize these properties $\Rightarrow$ we first need to develop new methods!

EL Embeddings

- Intelligent decisions need a "world model"
  - "know" facts that are true in the world
  - "infer" facts that are necessarily true
- can neural networks have a world model?
- "model-generating" embedding:
  - maps symbols into $\Re^n$ while preserving their model-theoretic semantics

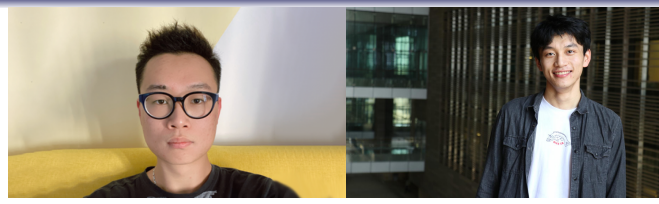
EL Embeddings



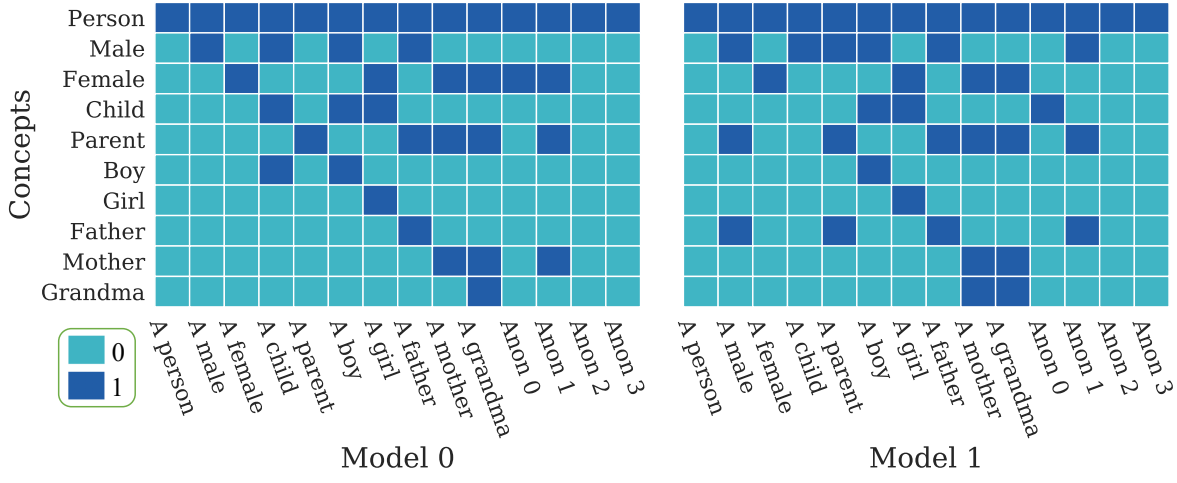
$$
\begin{aligned}
\mathit{Male} &\sqsubseteq \mathit{Person} \\
\mathit{Female} &\sqsubseteq \mathit{Person} \\
\mathit{Father} &\sqsubseteq \mathit{Male} \\
\mathit{Mother} &\sqsubseteq \mathit{Female} \\
\mathit{Father} &\sqsubseteq \mathit{Parent} \\
\mathit{Mother} &\sqsubseteq \mathit{Parent} \\
\mathit{Female} \sqcap \mathit{Male} &\sqsubseteq \bot \\
\mathit{Female} \sqcap \mathit{Parent} &\sqsubseteq \mathit{Mother} \\
\mathit{Male} \sqcap \mathit{Parent} &\sqsubseteq \mathit{Father} \\
\exists \mathit{hasChild}.\mathit{Person} &\sqsubseteq \mathit{Parent} \\
\mathit{Parent} &\sqsubseteq \mathit{Person} \\
\mathit{Parent} &\sqsubseteq \exists \mathit{hasChild}.\top
\end{aligned}
$$

Kulmanov et al., 2020. IJCAI

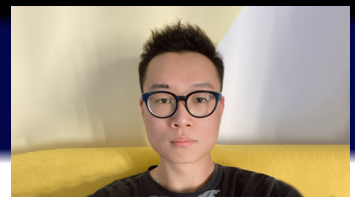
Single models are not enough for entailment



- distinguishing possibility and necessity
  - true in “some” worlds
  - true in “all” worlds



FALCON: Fuzzy generation of models for $\mathcal{ALC}$



Algorithm 1 Generating $C^{\mathcal{I}}$ for a Concept Description C.

Function Calculate $m(\cdot, C^{\mathcal{I}})$

Require: Embedding function f_e; Multilayer Perceptron MLP; Activation function σ; Sampling size k; Fuzzy operators θ, κ, ν; Individuals $\mathbf{I} = \mathbf{I}_n \cup \mathbf{I}_{\mathfrak{R}^n}$

Sample X with $|X| = k$ from $\mathbf{I}$
Compute $m(X, C^{\mathcal{I}}) := \{m(x, C^{\mathcal{I}}) | x \in X\}$:
if C is a concept name **then**
  $m(X, C^{\mathcal{I}}) = \sigma(MLP(f_e(C), f_e(X)))$
else if $C = C_1 \sqcap C_2$ **then**
  $m(X, (C_1 \sqcap C_2)^{\mathcal{I}}) = \theta(m(X, C_1^{\mathcal{I}}), m(x, C_2^{\mathcal{I}}))$
else if $C = C_1 \sqcup C_2$ **then**
  $m(X, (C_1 \sqcup C_2)^{\mathcal{I}}) = \kappa(m(X, C_1^{\mathcal{I}}), m(X, C_2^{\mathcal{I}}))$
else if $C = \neg D$ **then**
  $m(X, (\neg D)^{\mathcal{I}}) = \nu(m(X, D^{\mathcal{I}}))$
else if $C = \exists R.D$ **then**
  Sample Y with $|Y| = k$ from $\mathbf{I}$
  $m(X, (\exists R.D)^{\mathcal{I}}) = \max_{y \in Y} \theta(m(y, D^{\mathcal{I}}), m((X, y), R^{\mathcal{I}}))$
  with $m((x, y), R^{\mathcal{I}}) = \sigma(MLP(f_e(x) + f_e(R), f_e(y)))$
else if $C = \forall R.D$ **then**
  Sample Y with $|Y| = k$ from $\mathbf{I}$
  $m(X, (\forall R.D)^{\mathcal{I}}) = \min_{y \in Y} \kappa(\nu(m(y, D^{\mathcal{I}})), m((X, y), R^{\mathcal{I}}))$
  with $m((x, y), R^{\mathcal{I}}) = \sigma(MLP(f_e(x) + f_e(R), f_e(y)))$
end if
return $m(X, C^{\mathcal{I}})$

FALCON: Fuzzy generation of models for $\mathcal{ALC}$

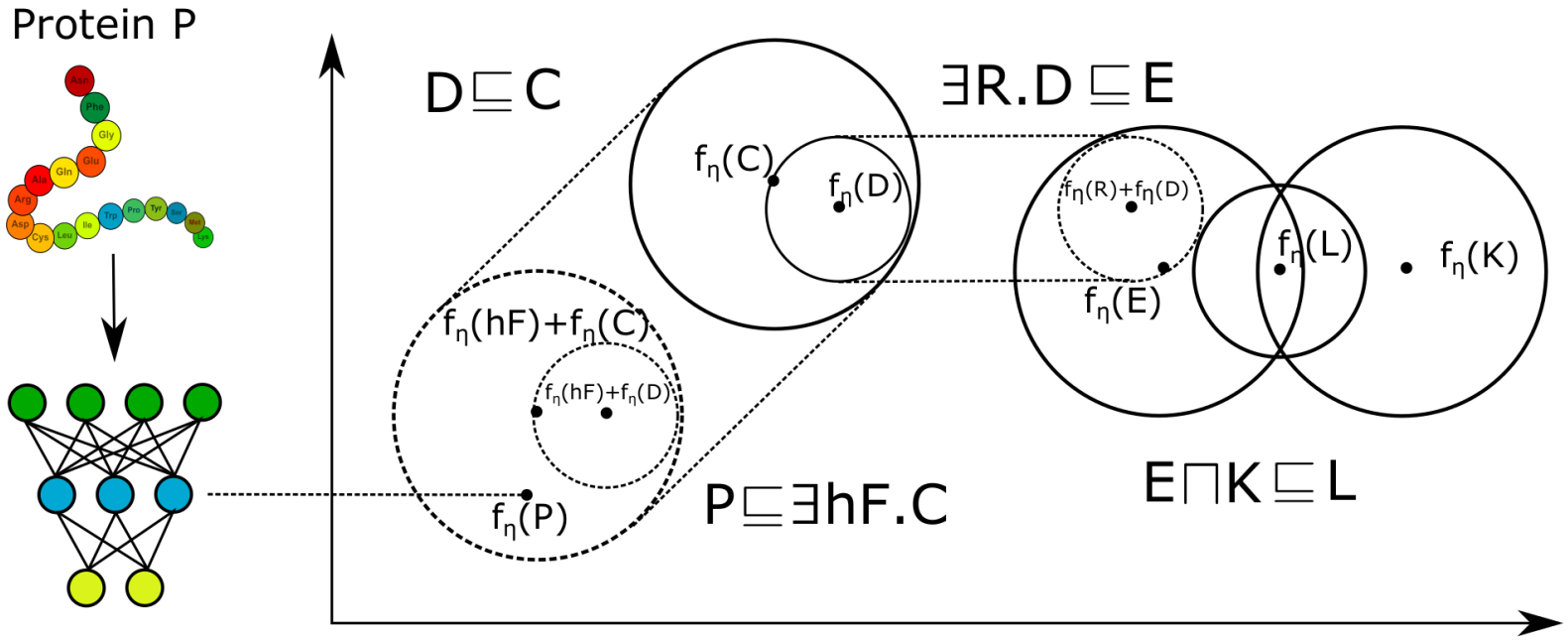
- Differentiable fuzzy logic to generate single models
  - using a recursive forward function to handle arbitrary concept descriptions
- sound and complete:
  - generates a model if and only if a model exists
- semantic entailment:
  - $T \models \phi$ iff $Mod(T) \subseteq Mod(\{\phi\})$
- enables:
  - reasoning under inconsistency, paraconsistent reasoning
  - combining prediction and deduction
  - knowledge-based zero-shot prediction

Tang et al., 2022

Where ontologies can help: little or no training data

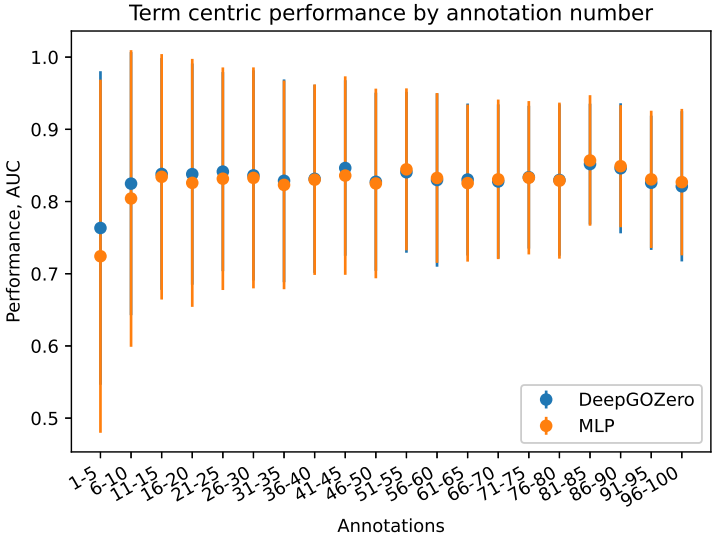
- some domains have little or no training data available:
  - metagenomic dark matter (orphan proteins)
  - rare diseases
  - genotype–phenotype relations in most populations
  - emerging pathogens
- will benefit from knowledge-enhanced predictions
  - prediction + inference
  - approximate inference

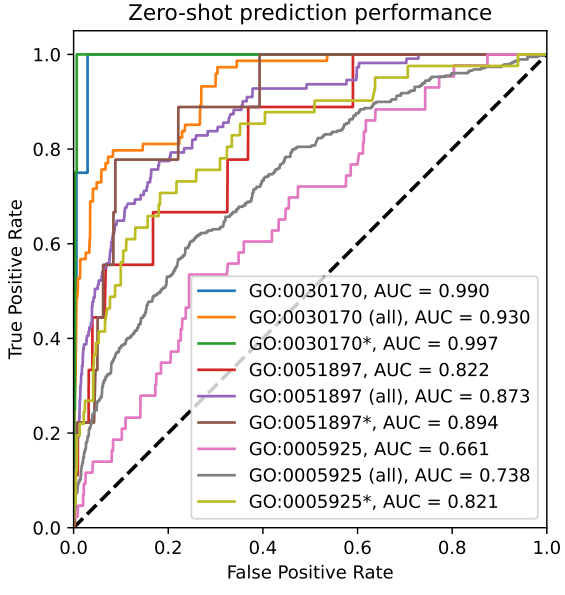
Zero-shot protein function prediction



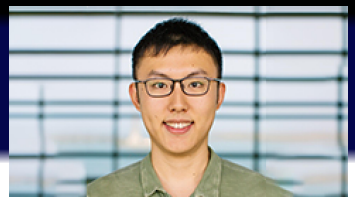
Kulmanov & Hoehndorf, DeepGOZero: Improving protein function prediction from sequence and zero-shot learning based on ontology axioms. ISMB, 2022.

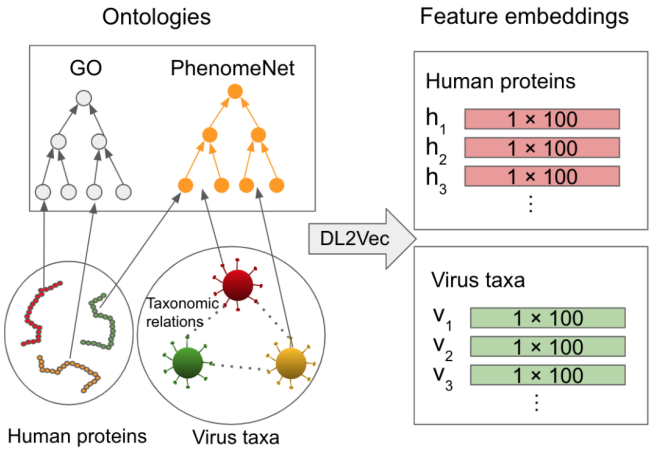
Zero shot protein function prediction



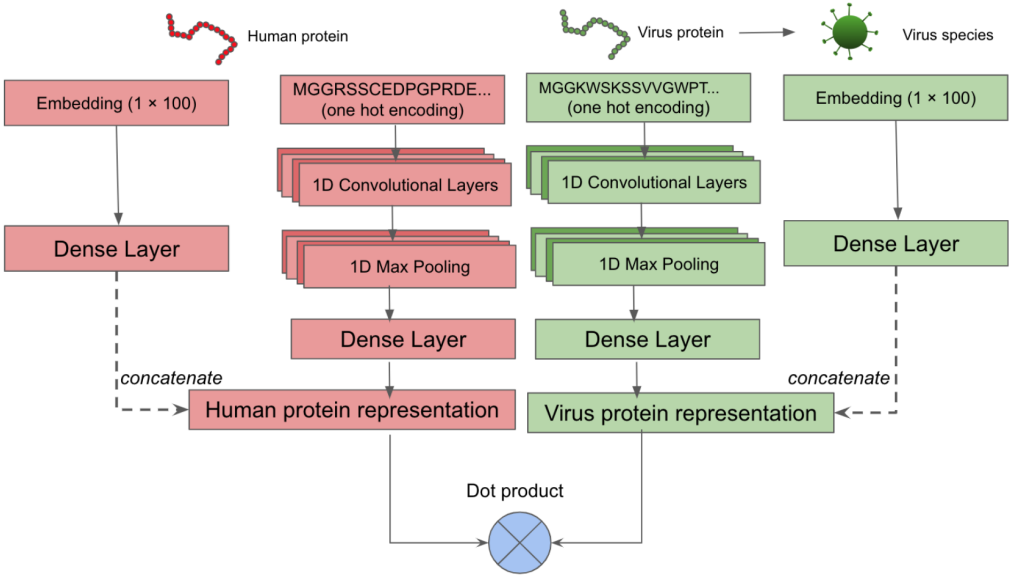


Injecting background knowledge: DeepViral



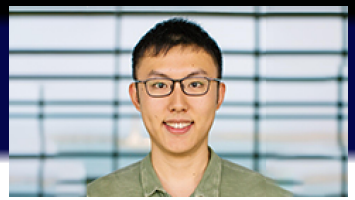


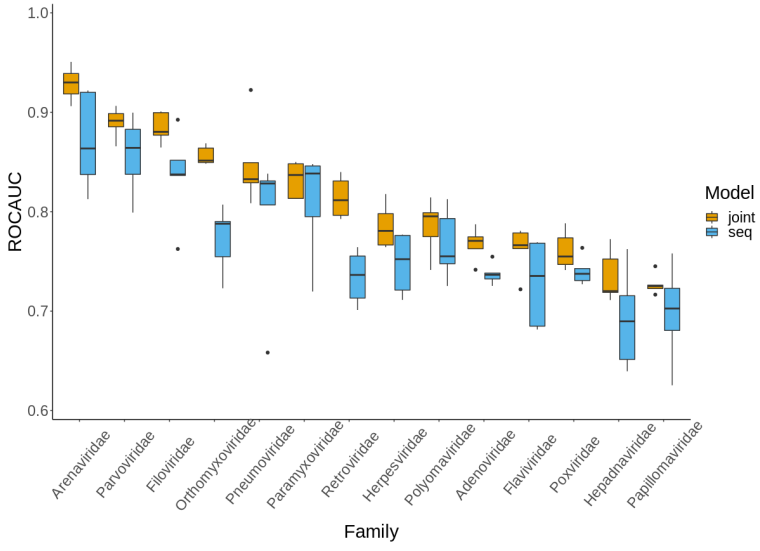
(a) Generation of feature embeddings

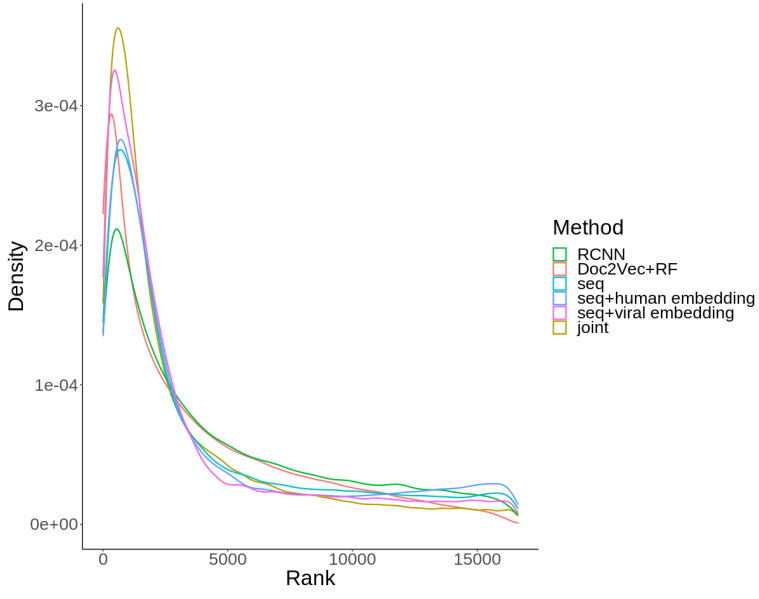


(b) Joint prediction model from embeddings and sequences

Injecting background knowledge: DeepViral

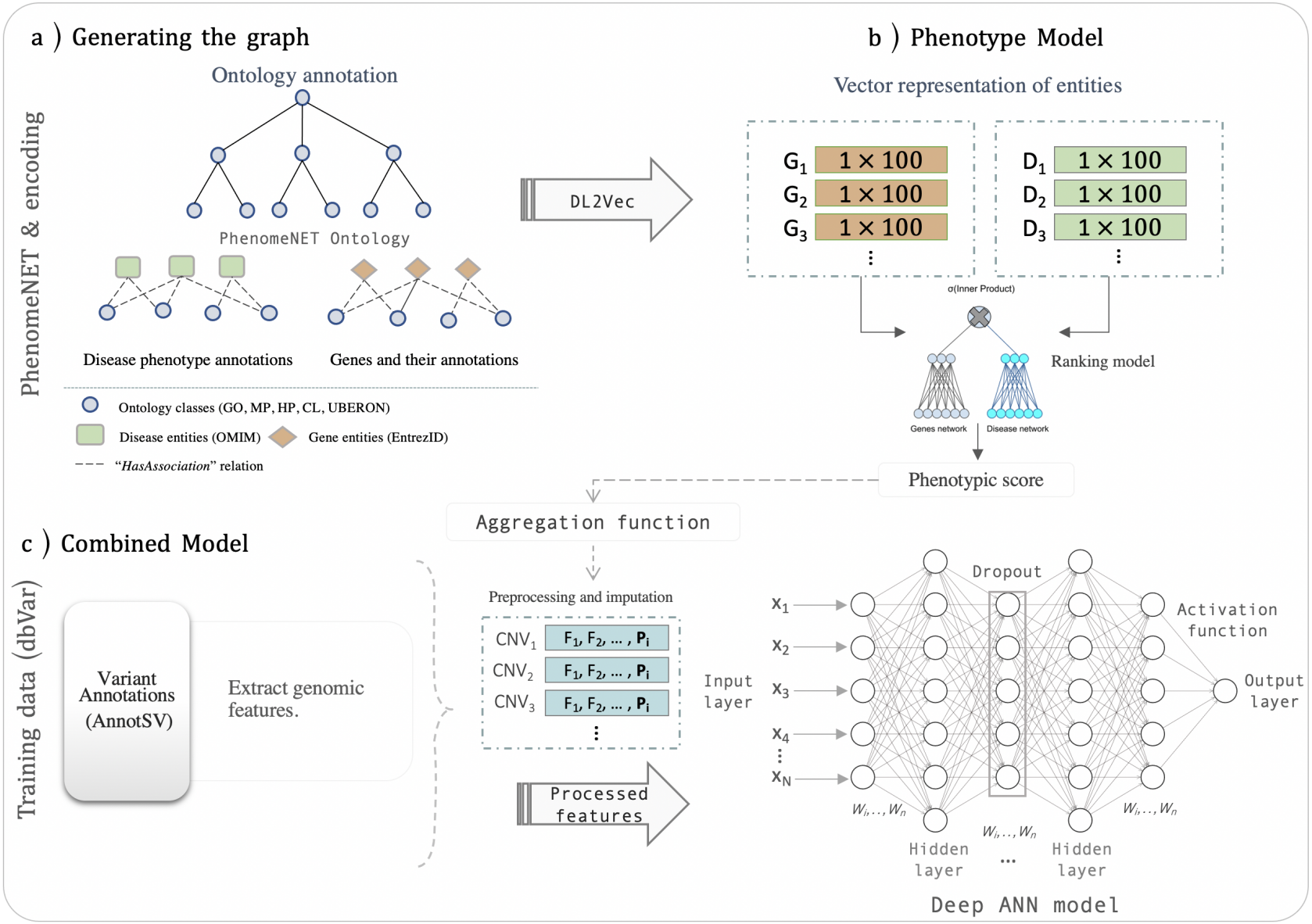






Wang et al., 2021

DeepSVP


a) Generating the graph
PhenomeNET & encoding
Ontology annotation
PhenomeNET Ontology
Disease phenotype annotations
Genes and their annotations
Ontology classes (GO, MP, HP, CL, UBERON)
Disease entities (OMIM)
Gene entities (EntrezID)
"HasAssociation" relation
DL2Vec
b) Phenotype Model
Vector representation of entities
G1 1 × 100
G2 1 × 100
G3 1 × 100
D1 1 × 100
D2 1 × 100
D3 1 × 100
σ(Inner Product)
Ranking model
Genes network
Disease network
Phenotypic score
Aggregation function
c) Combined Model
Training data (dbVar)
Variant Annotations (AnnotSV)
Extract genomic features.
Preprocessing and imputation
CNV1 F1, F2, ... , Pi
CNV2 F1, F2, ... , Pi
CNV3 F1, F2, ... , Pi
Processed features
Input layer
Dropout
Activation function
Output layer
Hidden layer
Hidden layer
Deep ANN model

DeepSVP

DeepSVP models using maximum score	**GO**	325 (0.2162)	536 (0.3566)	725 (0.4824)	0.9558	0.2670
	MP	237 (0.1577)	630 (0.4192)	855 (0.5689)	0.9605	0.2492
	HP	445 (0.2961)	**1088 (0.7239)**	**1348 (0.8969)**	**0.9929**	0.4364
	CL	272 (0.1810)	835 (0.5556)	1148 (0.7638)	0.9801	0.2569
	UBERON	259 (0.1723)	637 (0.4238)	1049 (0.6979)	0.9733	0.2417
	Union	328 (0.2182)	948 (0.6307)	1122 (0.7465)	0.9750	0.3489
SV pathogenicity prediction/ranking	**StrVCTVRE**	72 (0.0479)	223 (0.1484)	405 (0.2695)	0.9178	0.0952
	CADD-SV	38 (0.0253)	620 (0.4125)	1020 (0.6786)	0.9816	0.1262
	AnnotSV	19 (0.0126)	229 (0.1524)	700 (0.4657)	0.9605	0.2203

Althagafi et al., 2022

mOWL

- high-performance software library for machine learning with Semantic Web (OWL) ontologies
- ontology embeddings, zero-shot predictions, knowledge-enhanced predictions
- Algorithms written in Python + Scala (OWLAPI), tuned for performance
- full access to OWLAPI from Python

`https://github.com/bio-ontology-research-group/mowl`

Timeline

Biology/Bioinformatics

- First eukaryotes sequenced
- Microarray experiments
- High throughput experiments
- Clinical integration
- Machine learning and AI

Ontologies

- Gene Ontology
- (Bio-)Ontological foundations
- Dealing with Big Data
- Knowledge graphs
- ???

Timeline

Biology/Bioinformatics

- First eukaryotes sequenced
- Microarray experiments
- High throughput experiments
- Clinical integration
- Machine learning and AI

Ontologies

- Gene Ontology
- (Bio-)Ontological foundations
- Dealing with Big Data
- Knowledge graphs
- ???

Will we need new methods or will ontologies have to change?

Probably both.

Ontologies and machine learning

- exploiting axioms, removing incorrect axioms becomes more and more relevant
  - very relevant now (see Sarah's poster at the poster session)
- as long as ontologies capture only "relatedness", knowledge graphs and knowledge graph analytics will suffice
- negation and disjointness axioms are really useful for reducing search space
- ontologies should enable useful *deductive inference* $\Rightarrow$ not found in knowledge graphs
  - deduction is a hallmark of "intelligent" systems

Acknowledgements

Thank you

Amyloid beta
Protein classified with blood
coagulation.

A Semantic Haiku

generated from the UniProt Knowledgebase

`http://borg.kaust.edu.sa`
robert.hoehndorf@kaust.edu.sa