Relational data and knowledge

Embedding ontologies: approaches

- syntactic: treat axioms as "sentences" using language models
- graph-based: treat ontologies as graphs (like in semantic similarity)
- model-theoretic: encode model-theoretic semantics in optimization

Ontology embeddings

Definition

Let $O = (\Sigma = (C, R, I); ax; \vdash)$ be an ontology with a set of classes C, a set of relations R, a set of instances I, a set of axioms ax and an inference relation $\vdash$. An ontology embedding is a function $f_\eta : C \cup R \cup I \mapsto \mathbb{R}^n$ (or $\Sigma(O) \mapsto \mathbb{R}^n$ (subject to certain constraints).

Ontology embeddings

Definition

Let $O = (\Sigma = (C, R, I); ax; \vdash)$ be an ontology with a set of classes C, a set of relations R, a set of instances I, a set of axioms ax and an inference relation $\vdash$. An ontology embedding is a function $f_\eta : C \cup R \cup I \mapsto \mathbb{R}^n$ (or $\Sigma(O) \mapsto \mathbb{R}^n$ (subject to certain constraints).

For example, we can use co-occurrence within $ax^\vdash$ to constrain the embedding function, where the constraints on co-occurrence are formulated using the Word2Vec skipgram model.

Word2Vec

Maximize:

$$\frac{1}{N}\sum_{n=1}^{N}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{n+j}|w_n) \tag{1}$$

with

$$p(w_o|w_i) = \frac{\exp({v'_{w_o}}^T v_{w_i})}{\sum_{w=1}^{W}\exp({v'_w}^T v_{w_i})} \tag{2}$$

(at least conceptually; different strategies are used to approximate Eqn. 2)

How to measure similarity?

- Shortest Path
  - applicable to arbitrary “knowledge graphs”
  - does not capture similarity well over all edge types, e.g., *disjointWith*, *differentFrom*, *opposite-of*, etc.
- Random Walk
  - with or without restart
  - iterated
  - does not consider edge labels $\Rightarrow$ captures only adjacency of nodes
  - scores whole graph with *probability* of being in a state
  - can take multiple seed nodes
    - can be used to find disease genes

Graph-based learning

- feature learning on graphs

Graph-based learning

- feature learning on graphs
- e.g., iterated, edge-labeled random walk
  - walks form *sentences*
  - sentences form a *corpus*
  - feature learning on corpus through Word2Vec (or factorization of co-occurrence matrix)
  - RDF2Vec: `http://data.dws.informatik.uni-mannheim.de/rdf2vec/`
  - with support for reasoning over ontologies: `https://github.com/bio-ontology-research-group/walking-rdf-and-owl`

Graph-based learning

- feature learning on graphs
- e.g., iterated, edge-labeled random walk
  - walks form *sentences*
  - sentences form a *corpus*
  - feature learning on corpus through Word2Vec (or factorization of co-occurrence matrix)
  - RDF2Vec: `http://data.dws.informatik.uni-mannheim.de/rdf2vec/`
  - with support for reasoning over ontologies: `https://github.com/bio-ontology-research-group/walking-rdf-and-owl`
- Translational knowledge graph embeddings: TransE, TransR, TransE, HolE, etc.
  - analogy- or translation-based
  - `https://github.com/SmartDataAnalytics/PyKEEN`

Graph-based learning

- feature learning on graphs
- e.g., iterated, edge-labeled random walk
  - walks form *sentences*
  - sentences form a *corpus*
  - feature learning on corpus through Word2Vec (or factorization of co-occurrence matrix)
  - RDF2Vec: `http://data.dws.informatik.uni-mannheim.de/rdf2vec/`
  - with support for reasoning over ontologies: `https://github.com/bio-ontology-research-group/walking-rdf-and-owl`
- Translational knowledge graph embeddings: TransE, TransR, TransE, HolE, etc.
  - analogy- or translation-based
  - `https://github.com/SmartDataAnalytics/PyKEEN`
- Graph Convolution Neural Networks (not discussed here)

Graph embeddings

Definition

Let $KG = (V, E, L; \vdash)$ be an ontology graph with a set of vertices V, a set of edges $E \subseteq V \times V$, a label function $L : V \cup E \mapsto Lab$ that assigns labels from a set of labels Lab to vertices and edges, and an inference relation $\vdash$. An ontology graph embedding is a function $f_\eta : L(V) \cup L(E) \mapsto \mathbf{R}^n$.

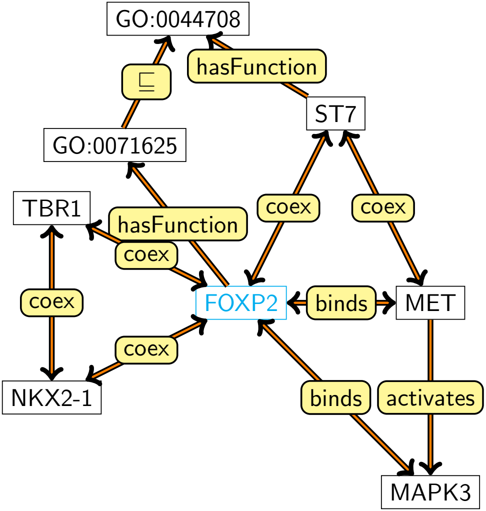
Graph embeddings

Definition

Let $KG = (V, E, L; \vdash)$ be an ontology graph with a set of vertices V, a set of edges $E \subseteq V \times V$, a label function $L : V \cup E \mapsto Lab$ that assigns labels from a set of labels Lab to vertices and edges, and an inference relation $\vdash$. An ontology graph embedding is a function $f_\eta : L(V) \cup L(E) \mapsto \mathbf{R}^n$.

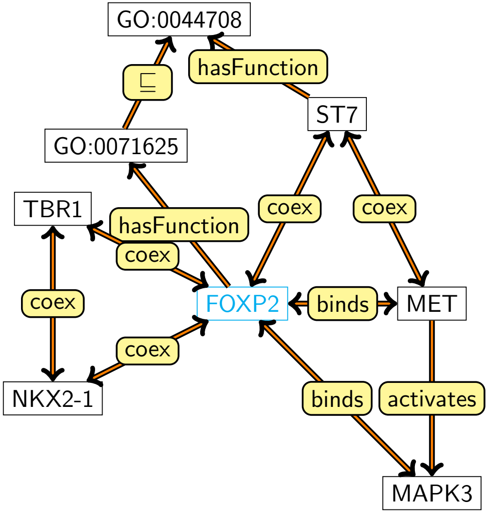
- key idea: preserve *some* structure of the graph in $\mathbb{R}^n$ (under operations in $\mathbb{R}^n$)
- $\mathbb{R}^n$ enables *new* operations (such as many similarity measures)
- useful as *feature* vectors

Random walks



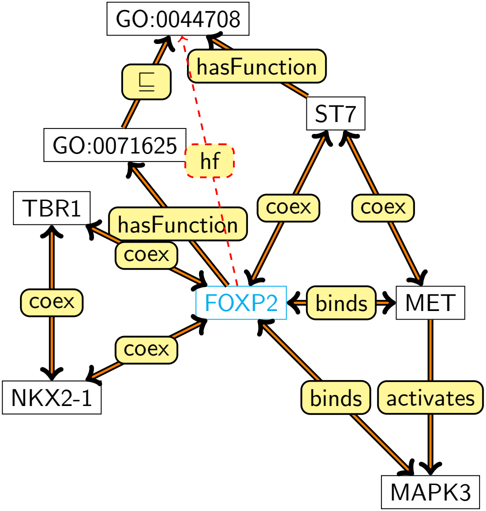
- FOXP2 is characterized by *adjacent* and close nodes and edges
- different edges may "transmit" information differently

Random walks



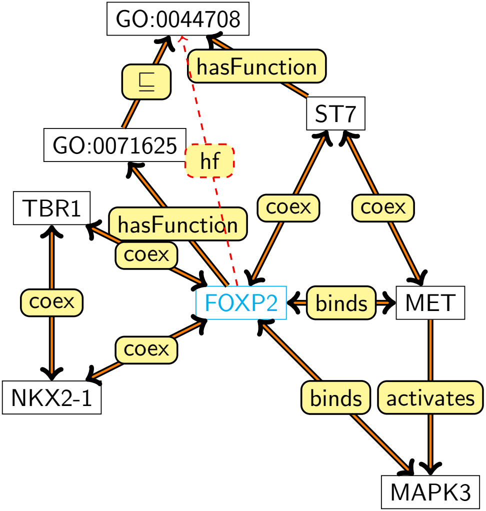
- precompute the deductive closure:
- for all ϕ: if $\mathcal{KG} \models \phi$, add ϕ to $\mathcal{KG}$

Random walks



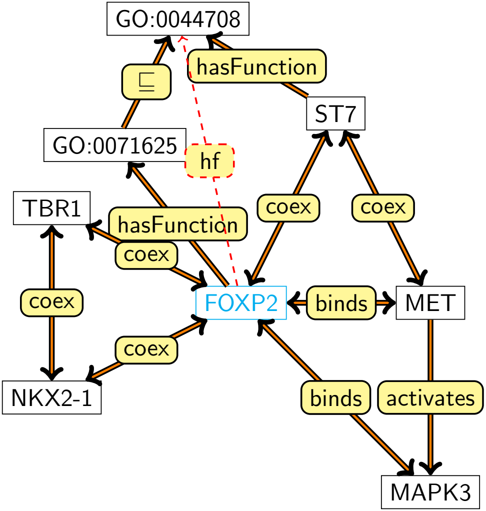
- precompute the deductive closure:
- for all ϕ: if $\mathcal{KG} \models \phi$, add ϕ to $\mathcal{KG}$

Random walks



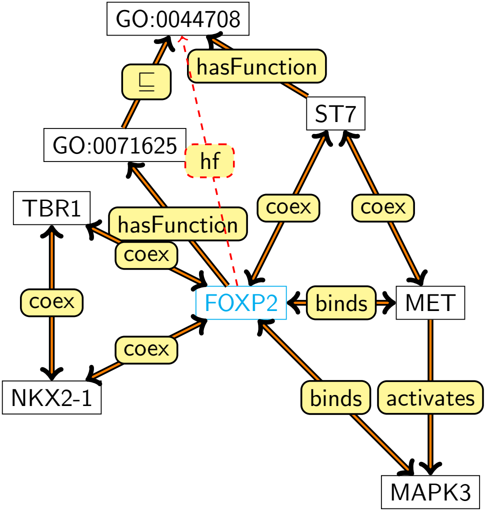
- Exploring the graph:

Random walks



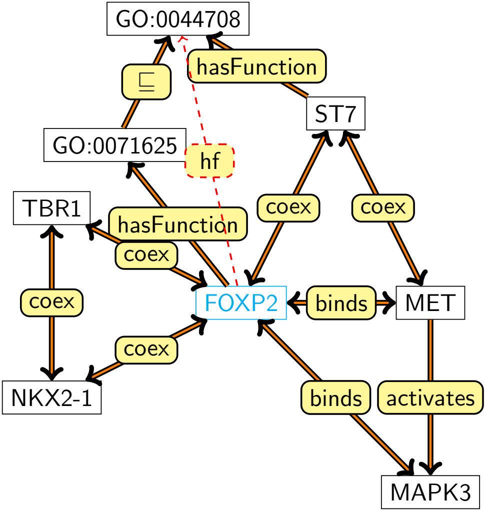
- Exploring the graph:
- :FOXP2 :binds :MET :coex :ST7 :hasFunction GO:0044708

Random walks



- Exploring the graph:
- :FOXP2 :binds :MET :coex :ST7 :hasFunction GO:0044708
- :FOXP2 :hasFunction GO:0071625 subClassOf GO:0044708

Random walks



- Exploring the graph:
- :FOXP2 :binds :MET :coex :ST7 :hasFunction GO:0044708
- :FOXP2 :hasFunction GO:0071625 subClassOf GO:0044708
- :FOXP2 :coex :TBR1 :coex :NKX2-1 :coex :TBR1 :coex ...

Word2Vec and Random Walks

- random walks "flatten" a graph
  - walks capture node neighborhood
  - and generate a "corpus"
- random walks capture graph "structure"
  - in ABox and TBox
  - hub-nodes, communities, etc.
  - determine "importance" of nodes
- embeddings capture co-occurrence
  - similar graph neighborhood $\Rightarrow$ similar co-occurrence $\Rightarrow$ similar vector
- embeddings generate "feature" vectors
  - functions from symbols (words, labels) into $\mathbb{R}^n$

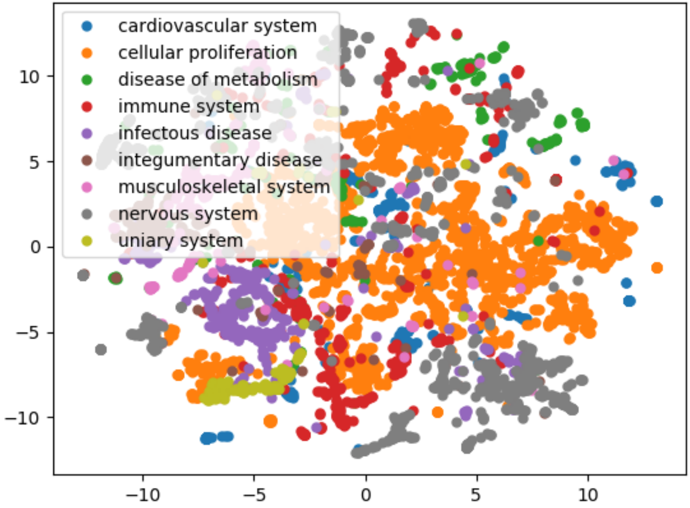
What to do with embeddings?

- useful for edge prediction, similarity, clustering, as feature vectors
  - supervised: edge prediction (e.g., SVM, ANN)
    - e.g.: find a function $f : \mathbb{R}^n \times \mathbb{R}^n \mapsto [0,1]$ s.t. $\sqrt{\frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}{T}}$ (RMSE) is minimized for a set of true labels y_k
  - unsupervised: clustering, similarity, visualization
    - cosine similarity (for L2-normalized features)
    - Word2Vec embeddings capture similarity between co-occurrence vectors

Visualizing feature vectors: dimensionality reduction

- project n-dimensional vectors in 2D (or 3D) space
- and color with some known labels
  - high-level/general classes in an ontology work great
- PCA or t-SNE
- `https://lvdmaaten.github.io/tsne/`

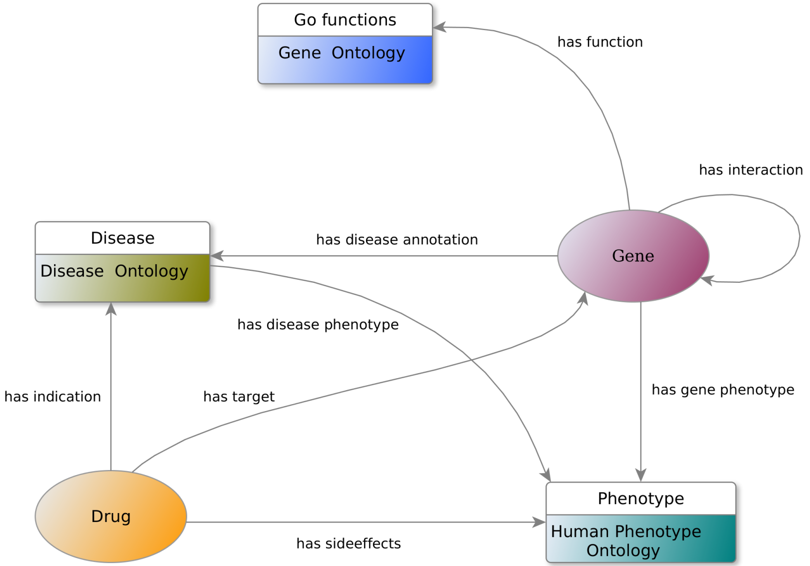
Visualizing feature vectors



Features: supervised learning

- feature vectors represent graph neighborhood of nodes
  - adjacent nodes and edges
  - ontology classes (asserted & inferred)
- useful in supervised prediction tasks
- relation prediction:
  - input: two features vectors (from embedding function)
  - output: 0 or 1 (relation or not)
  - training data: positive and negative cases
    - $R(x, y)$ and $\neg R(x, y)$
    - $R(x, y)$ and not provable $R(x, y)$

Features: supervised learning



Features: supervised learning

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, graphs, and text

The forkhead-box P2 (FOXP2) gene polymorphism has been reported to be involved in the susceptibility to schizophrenia; however, few studies have investigated the association between FOXP2 gene polymorphism and clinical symptoms in schizophrenia.

Ontologies, graphs, and text

The forkhead-box P2 (FOXP2) gene polymorphism has been reported to be involved in the susceptibility to schizophrenia; however, few studies have investigated the association between FOXP2 gene polymorphism and clinical symptoms in schizophrenia.

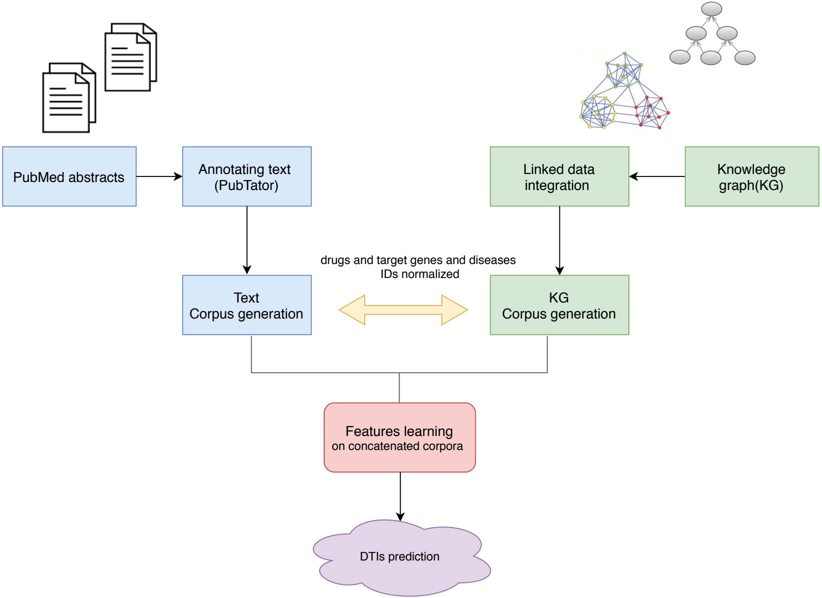
- :FOXP2 :binds :MET :coex :ST7 :hasFunction GO:0044708
- :FOXP2 :hasFunction GO:0071625 subClassOf GO:0044708
- :FOXP2 :coex :TBR1 :coex :NKX2-1 :coex :TBR1 :coex ...

Ontologies, graphs, and text

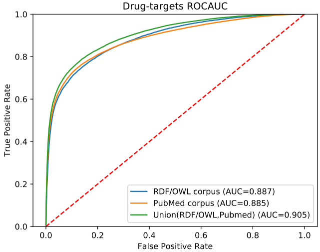
The :FOXP2 gene polymorphism has been reported to be involved in the susceptibility to schizophrenia; however, few studies have investigated the association between :FOXP2 gene polymorphism and clinical symptoms in schizophrenia.

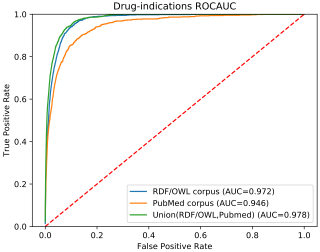
- :FOXP2 :binds :MET :coex :ST7 :hasFunction GO:0044708
- :FOXP2 :hasFunction GO:0071625 subClassOf GO:0044708
- :FOXP2 :coex :TBR1 :coex :NKX2-1 :coex :TBR1 :coex ...

Multi-modal feature learning



Multi-modal feature learning: drug targets and indications





Alshahrani & H. Drug repurposing through multi-modal learning on knowledge graphs. BioRxiv, 2018.

Tools and resources

- RDF2Vec: random walks on RDF + Word2Vec
- RDF2Vec: Weisfeiler-Lehmann kernel on RDF
- `https://datalab.rwth-aachen.de/embedding/RDF2Vec/`

Tools and resources

- RDF2Vec: random walks on RDF + Word2Vec
- RDF2Vec: Weisfeiler-Lehmann kernel on RDF
- `https://datalab.rwth-aachen.de/embedding/RDF2Vec/`
- Walking RDF+OWL: random walks on RDF + Elk + Word2Vec
  - inference
- `https://github.com/bio-ontology-research-group/walking-rdf-and-owl`

Some limitations

- "word"-based (Word2Vec):
  - semantics is reduced to co-occurrence (in ABox/TBox statements)
  - "disjointWith" vs. "part-of" vs. "subClassOf"

Translating embeddings

Definition

Let $KG = (V, E, L; \vdash)$ be a knowledge graph with a set of vertices V, a set of edges $E \subseteq V \times V$, a label function $L : V \cup E \mapsto Lab$ that assigns labels from a set of labels Lab to vertices and edges, and an inference relation $\vdash$. A knowledge graph embedding is a function $f_\eta : L(V) \cup L(E) \mapsto \mathbf{R}^n$.

Translating embeddings

Definition

Let $KG = (V, E, L; \vdash)$ be a knowledge graph with a set of vertices V, a set of edges $E \subseteq V \times V$, a label function $L : V \cup E \mapsto Lab$ that assigns labels from a set of labels Lab to vertices and edges, and an inference relation $\vdash$. A knowledge graph embedding is a function $f_\eta : L(V) \cup L(E) \mapsto \mathbf{R}^n$.

Graph as edgelist: set of (s, p, o) statements

Translating embeddings

Definition

Let $KG = (V, E, L; \vdash)$ be a knowledge graph with a set of vertices V, a set of edges $E \subseteq V \times V$, a label function $L : V \cup E \mapsto Lab$ that assigns labels from a set of labels Lab to vertices and edges, and an inference relation $\vdash$. A knowledge graph embedding is a function $f_\eta : L(V) \cup L(E) \mapsto \mathbf{R}^n$.

Graph as edgelist: set of (s, p, o) statements
Idea: $\mu(s) + \mu(p) \approx \mu(o)$

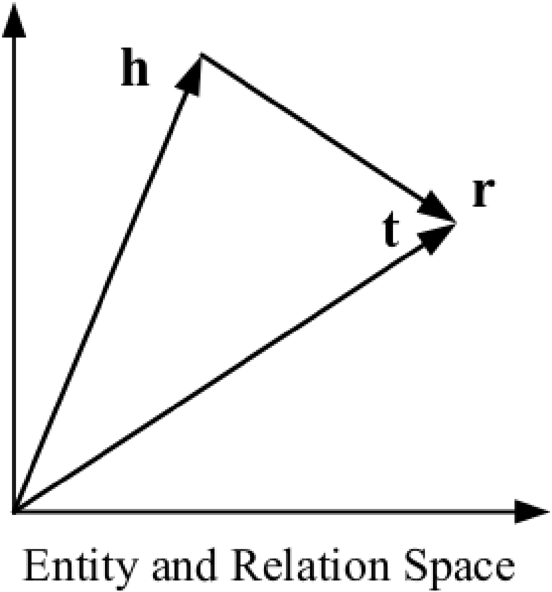
Translating embeddings

Definition

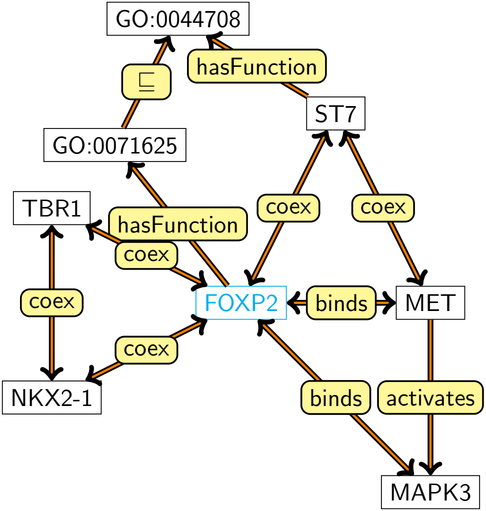
Let $KG = (V, E, L; \vdash)$ be a knowledge graph with a set of vertices V, a set of edges $E \subseteq V \times V$, a label function $L : V \cup E \mapsto Lab$ that assigns labels from a set of labels Lab to vertices and edges, and an inference relation $\vdash$. A knowledge graph embedding is a function $f_\eta : L(V) \cup L(E) \mapsto \mathbf{R}^n$.

Graph as edgelist: set of (s, p, o) statements
Idea: $\mu(s) + \mu(p) \approx \mu(o)$
Minimize: $\sum_t \|\mu(s) + \mu(p) - \mu(o)\|$ (chose your norm, usually L2)

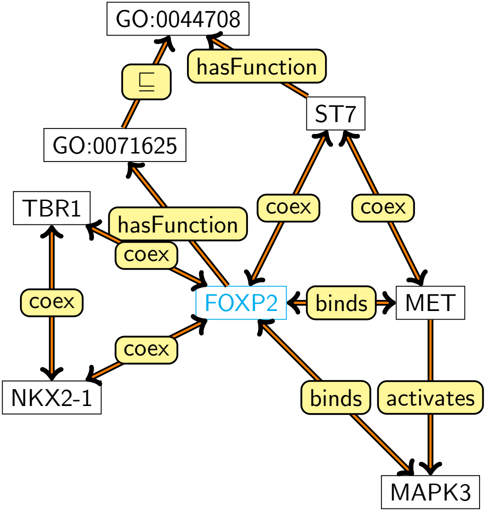
Translating embeddings



Translating embeddings

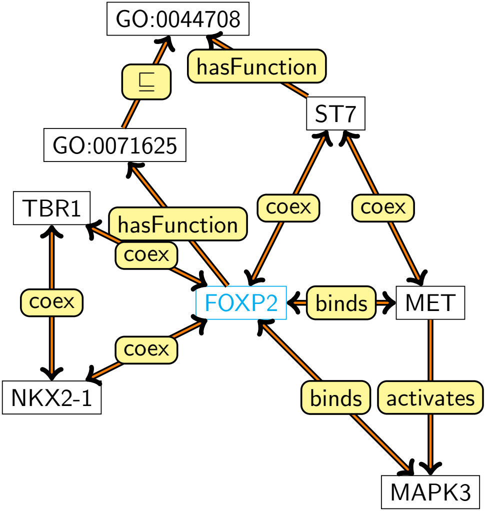


Translating embeddings



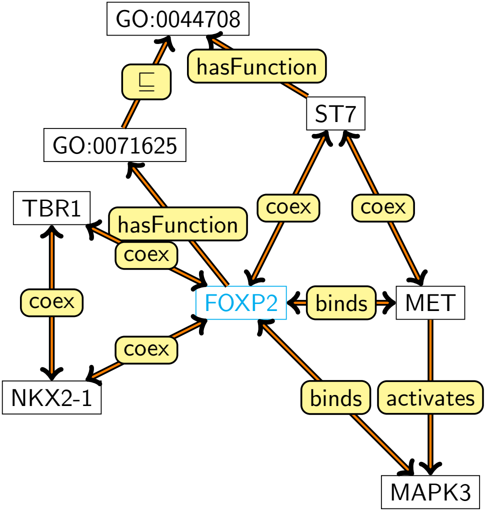
- FOXP2 + binds = MET

Translating embeddings



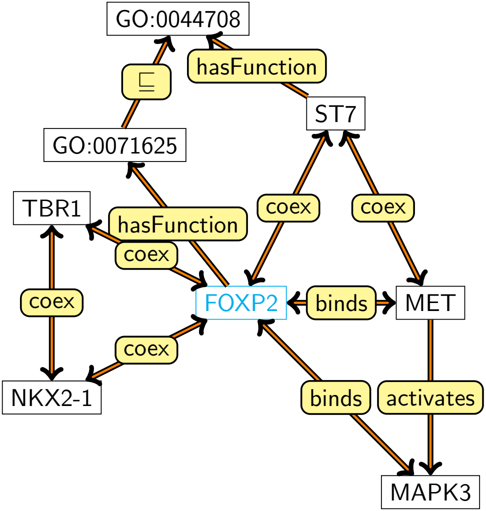
- FOXP2 + binds = MET
- MET + activates = MAPK3

Translating embeddings



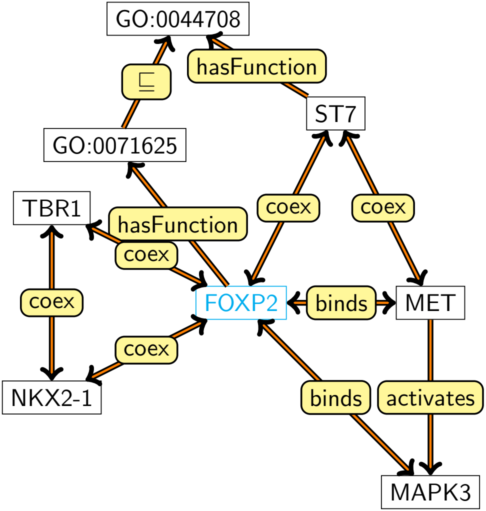
- FOXP2 + binds = MET
- MET + activates = MAPK3
- MET + binds = FOXP2

Translating embeddings



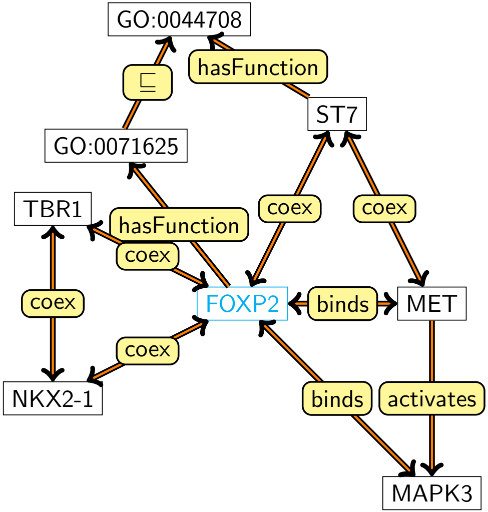
- FOXP2 + binds = MET
- MET + activates = MAPK3
- MET + binds = FOXP2
- ST7 + hasFunction = `GO:0044708`

Translating embeddings



- FOXP2 + binds = MET
- MET + activates = MAPK3
- MET + binds = FOXP2
- ST7 + hasFunction = `GO:0044708`
- ...

Translating embeddings



- FOXP2 + binds - MET = 0
- MAP + activates - MAPK3 = 0
- MET + binds - FOXP2 = 0
- ST7 + hasFunction - `GO:0044708` = 0
- ...

Translating embeddings

Algorithm 1 Learning TransE

input Training set $S = \{(h, \ell, t)\}$, entities and rel. sets E and L, margin γ, embeddings dim. k.

initialize $\boldsymbol{\ell} \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each $\ell \in L$

$\quad\quad\boldsymbol{\ell} \leftarrow \boldsymbol{\ell} / \|\boldsymbol{\ell}\|$ for each $\ell \in L$

$\quad\quad\mathbf{e} \leftarrow \text{uniform}(-\frac{6}{\sqrt{k}}, \frac{6}{\sqrt{k}})$ for each entity $e \in E$

loop

$\quad\mathbf{e} \leftarrow \mathbf{e} / \|\mathbf{e}\|$ for each entity $e \in E$

$\quad S_{batch} \leftarrow \text{sample}(S, b)$ // sample a minibatch of size b

$\quad T_{batch} \leftarrow \emptyset$ // initialize the set of pairs of triplets

$\quad$**for** $(h, \ell, t) \in S_{batch}$ **do**

$\quad\quad(h', \ell, t') \leftarrow \text{sample}(S'_{(h,\ell,t)})$ // sample a corrupted triplet

$\quad\quad T_{batch} \leftarrow T_{batch} \cup \{((h, \ell, t), (h', \ell, t'))\}$

$\quad$**end for**

$\quad$Update embeddings w.r.t. $\displaystyle\sum_{((h,\ell,t),(h',\ell,t')) \in T_{batch}} \nabla\left[\gamma + d(\boldsymbol{h} + \boldsymbol{\ell}, \boldsymbol{t}) - d(\boldsymbol{h'} + \boldsymbol{\ell}, \boldsymbol{t'})\right]_+$

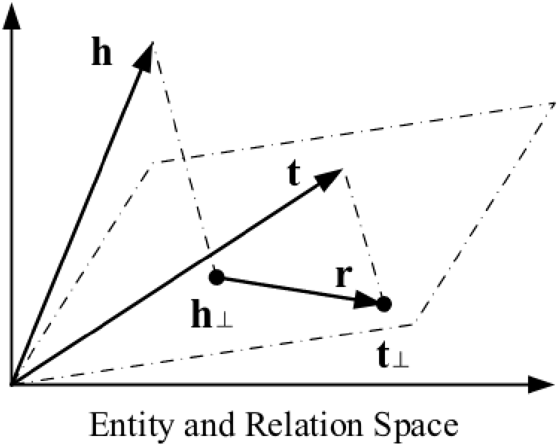
end loop

Bordes et al. (2013). Translating Embeddings for Modeling Multi-relational Data.

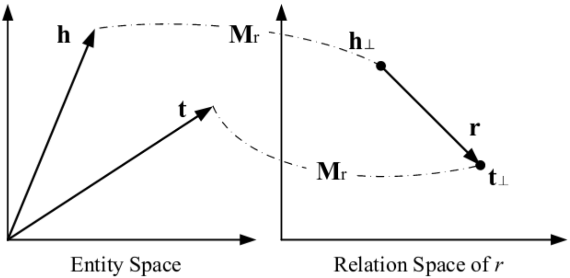
Some properties of TransE

- graph-based
  - works well on RDF graphs
  - and ontology graphs
- 1:1 relations only
  - not suitable for hierarchies (1-N relations)
  - not suitable for N-N relations
  - no transitive, symmetric, reflexive relations

Translating embeddings



Translating embeddings



(c) TransR.

Translating embeddings

Method	Ent. embedding	Rel. embedding	Scoring function $f_r(h,t)$	Constraints/Regularization
TransE [14]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{r}\in\mathbb{R}^d$	$-\lVert\mathbf{h}+\mathbf{r}-\mathbf{t}\rVert_{1/2}$	$\lVert\mathbf{h}\rVert_2=1,\lVert\mathbf{t}\rVert_2=1$
TransH [15]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{r},\mathbf{w}_r\in\mathbb{R}^d$	$-\lVert(\mathbf{h}-\mathbf{w}_r^\top\mathbf{h}\mathbf{w}_r)+\mathbf{r}-(\mathbf{t}-\mathbf{w}_r^\top\mathbf{t}\mathbf{w}_r)\rVert_2^2$	$\lVert\mathbf{h}\rVert_2\le 1,\lVert\mathbf{t}\rVert_2\le 1$ $\lvert\mathbf{w}_r^\top\mathbf{r}\rvert/\lVert\mathbf{r}\rVert_2\le\epsilon,\lVert\mathbf{w}_r\rVert_2=1$
TransR [16]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{r}\in\mathbb{R}^k,\mathbf{M}_r\in\mathbb{R}^{k\times d}$	$-\lVert\mathbf{M}_r\mathbf{h}+\mathbf{r}-\mathbf{M}_r\mathbf{t}\rVert_2^2$	$\lVert\mathbf{h}\rVert_2\le 1,\lVert\mathbf{t}\rVert_2\le 1,\lVert\mathbf{r}\rVert_2\le 1$ $\lVert\mathbf{M}_r\mathbf{h}\rVert_2\le 1,\lVert\mathbf{M}_r\mathbf{t}\rVert_2\le 1$
TransD [50]	$\mathbf{h},\mathbf{w}_h\in\mathbb{R}^d$ $\mathbf{t},\mathbf{w}_t\in\mathbb{R}^d$	$\mathbf{r},\mathbf{w}_r\in\mathbb{R}^k$	$-\lVert(\mathbf{w}_r\mathbf{w}_h^\top+\mathbf{I})\mathbf{h}+\mathbf{r}-(\mathbf{w}_r\mathbf{w}_t^\top+\mathbf{I})\mathbf{t}\rVert_2^2$	$\lVert\mathbf{h}\rVert_2\le 1,\lVert\mathbf{t}\rVert_2\le 1,\lVert\mathbf{r}\rVert_2\le 1$ $\lVert(\mathbf{w}_r\mathbf{w}_h^\top+\mathbf{I})\mathbf{h}\rVert_2\le 1$ $\lVert(\mathbf{w}_r\mathbf{w}_t^\top+\mathbf{I})\mathbf{t}\rVert_2\le 1$
TranSparse [51]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{r}\in\mathbb{R}^k,\mathbf{M}_r(\theta_r)\in\mathbb{R}^{k\times d}$ $\mathbf{M}_r^1(\theta_r^1),\mathbf{M}_r^2(\theta_r^2)\in\mathbb{R}^{k\times d}$	$-\lVert\mathbf{M}_r(\theta_r)\mathbf{h}+\mathbf{r}-\mathbf{M}_r(\theta_r)\mathbf{t}\rVert_{1/2}^2$ $-\lVert\mathbf{M}_r^1(\theta_r^1)\mathbf{h}+\mathbf{r}-\mathbf{M}_r^2(\theta_r^2)\mathbf{t}\rVert_{1/2}^2$	$\lVert\mathbf{h}\rVert_2\le 1,\lVert\mathbf{t}\rVert_2\le 1,\lVert\mathbf{r}\rVert_2\le 1$ $\lVert\mathbf{M}_r(\theta_r)\mathbf{h}\rVert_2\le 1,\lVert\mathbf{M}_r(\theta_r)\mathbf{t}\rVert_2\le 1$ $\lVert\mathbf{M}_r^1(\theta_r^1)\mathbf{h}\rVert_2\le 1,\lVert\mathbf{M}_r^2(\theta_r^2)\mathbf{t}\rVert_2\le 1$
TransM [52]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{r}\in\mathbb{R}^d$	$-\theta_r\lVert\mathbf{h}+\mathbf{r}-\mathbf{t}\rVert_{1/2}$	$\lVert\mathbf{h}\rVert_2=1,\lVert\mathbf{t}\rVert_2=1$
ManifoldE [53]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{r}\in\mathbb{R}^d$	$-(\lVert\mathbf{h}+\mathbf{r}-\mathbf{t}\rVert_2^2-\theta_r^2)^2$	$\lVert\mathbf{h}\rVert_2\le 1,\lVert\mathbf{t}\rVert_2\le 1,\lVert\mathbf{r}\rVert_2\le 1$
TransF [54]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{r}\in\mathbb{R}^d$	$(\mathbf{h}+\mathbf{r})^\top\mathbf{t}+(\mathbf{t}-\mathbf{r})^\top\mathbf{h}$	$\lVert\mathbf{h}\rVert_2\le 1,\lVert\mathbf{t}\rVert_2\le 1,\lVert\mathbf{r}\rVert_2\le 1$
TransA [55]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{r}\in\mathbb{R}^d,\mathbf{M}_r\in\mathbb{R}^{d\times d}$	$-(\lvert\mathbf{h}+\mathbf{r}-\mathbf{t}\rvert)^\top\mathbf{M}_r(\lvert\mathbf{h}+\mathbf{r}-\mathbf{t}\rvert)$	$\lVert\mathbf{h}\rVert_2\le 1,\lVert\mathbf{t}\rVert_2\le 1,\lVert\mathbf{r}\rVert_2\le 1$ $\lVert\mathbf{M}_r\rVert_F\le 1,[\mathbf{M}_r]_{ij}=[\mathbf{M}_r]_{ji}\ge 0$
KG2E [45]	$\mathbf{h}\sim\mathcal{N}(\boldsymbol{\mu}_h,\boldsymbol{\Sigma}_h)$ $\mathbf{t}\sim\mathcal{N}(\boldsymbol{\mu}_t,\boldsymbol{\Sigma}_t)$ $\boldsymbol{\mu}_h,\boldsymbol{\mu}_t\in\mathbb{R}^d$ $\boldsymbol{\Sigma}_h,\boldsymbol{\Sigma}_t\in\mathbb{R}^{d\times d}$	$\mathbf{r}\sim\mathcal{N}(\boldsymbol{\mu}_r,\boldsymbol{\Sigma}_r)$ $\boldsymbol{\mu}_r\in\mathbb{R}^d,\boldsymbol{\Sigma}_r\in\mathbb{R}^{d\times d}$	$-\mathrm{tr}(\boldsymbol{\Sigma}_r^{-1}(\boldsymbol{\Sigma}_h+\boldsymbol{\Sigma}_t))-\boldsymbol{\mu}^\top\boldsymbol{\Sigma}_r^{-1}\boldsymbol{\mu}-\ln\frac{\det(\boldsymbol{\Sigma}_r)}{\det(\boldsymbol{\Sigma}_h+\boldsymbol{\Sigma}_t)}$ $-\boldsymbol{\mu}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}-\ln(\det(\boldsymbol{\Sigma}))$ $\boldsymbol{\mu}=\boldsymbol{\mu}_h+\boldsymbol{\mu}_r-\boldsymbol{\mu}_t$ $\boldsymbol{\Sigma}=\boldsymbol{\Sigma}_h+\boldsymbol{\Sigma}_r+\boldsymbol{\Sigma}_t$	$\lVert\boldsymbol{\mu}_h\rVert_2\le 1,\lVert\boldsymbol{\mu}_t\rVert_2\le 1,\lVert\boldsymbol{\mu}_r\rVert_2\le 1$ $c_{min}\mathbf{I}\le\boldsymbol{\Sigma}_h\le c_{max}\mathbf{I}$ $c_{min}\mathbf{I}\le\boldsymbol{\Sigma}_t\le c_{max}\mathbf{I}$ $c_{min}\mathbf{I}\le\boldsymbol{\Sigma}_r\le c_{max}\mathbf{I}$
TransG [46]	$\mathbf{h}\sim\mathcal{N}(\boldsymbol{\mu}_h,\sigma_h^2\mathbf{I})$ $\mathbf{t}\sim\mathcal{N}(\boldsymbol{\mu}_t,\sigma_t^2\mathbf{I})$ $\boldsymbol{\mu}_h,\boldsymbol{\mu}_t\in\mathbb{R}^d$	$\boldsymbol{\mu}_r^i\sim\mathcal{N}(\boldsymbol{\mu}_t-\boldsymbol{\mu}_h,(\sigma_h^2+\sigma_t^2)\mathbf{I})$ $\mathbf{r}=\sum_i\pi_r^i\boldsymbol{\mu}_r^i\in\mathbb{R}^d$	$\sum_i\pi_r^i\exp\left(-\frac{\lVert\boldsymbol{\mu}_h+\boldsymbol{\mu}_r^i-\boldsymbol{\mu}_t\rVert_2^2}{\sigma_h^2+\sigma_t^2}\right)$	$\lVert\boldsymbol{\mu}_h\rVert_2\le 1,\lVert\boldsymbol{\mu}_t\rVert_2\le 1,\lVert\boldsymbol{\mu}_r^i\rVert_2\le 1$
UM [56]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	—	$-\lVert\mathbf{h}-\mathbf{t}\rVert_2^2$	$\lVert\mathbf{h}\rVert_2=1,\lVert\mathbf{t}\rVert_2=1$
SE [57]	$\mathbf{h},\mathbf{t}\in\mathbb{R}^d$	$\mathbf{M}_r^1,\mathbf{M}_r^2\in\mathbb{R}^{d\times d}$	$-\lVert\mathbf{M}_r^1\mathbf{h}-\mathbf{M}_r^2\mathbf{t}\rVert_1$	$\lVert\mathbf{h}\rVert_2=1,\lVert\mathbf{t}\rVert_2=1$

Wang et al. Knowledge Graph Embedding: A Survey ofApproaches and Applications.

PyKEEN

- Python package to generate knowledge graph embeddings
- supports many different graph embedding types: TransE, TransR, TransD, RESCAL, etc.
- hyperparameter optimization ("HPO") and evaluation included
- `https://github.com/SmartDataAnalytics/PyKEEN`

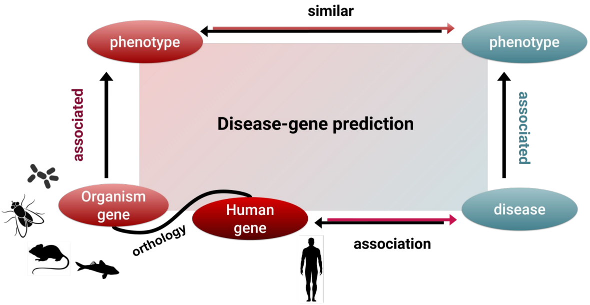
Applications of Ontology Embeddings

- Predicting gene-disease associations based on phenotypic similarity
- Diagnosis of disease based on phenotypic similarity
- Predict protein–protein interactions based on their functional similarity

….

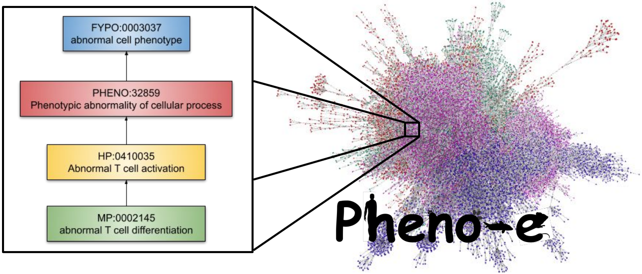
Predicting gene-disease associations

- Based on Phenotypic similarity
- Using the phenotypes of model organism genes and the diseases' phenotypes

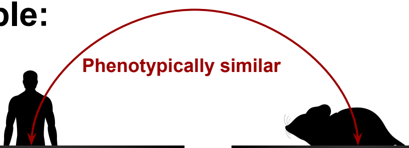


Predicting gene-disease associations

- PhenomeNet-Extension **(Pheno-e)** and **uPheno** are cross-species phenotype ontologies that can be utilized here.
- Both with the objective of allowing similar phenotypes from the same or different organisms ontologies to be logically defined in similar form.



Example:



Isolated anhidrosis with normal morphology and number sweat glands (ANHD)

Itpr2
inositol 1,4,5-triphosphate receptor 2

Human disease phenotypes:

Generalized anhidrosis

Heat intolerance

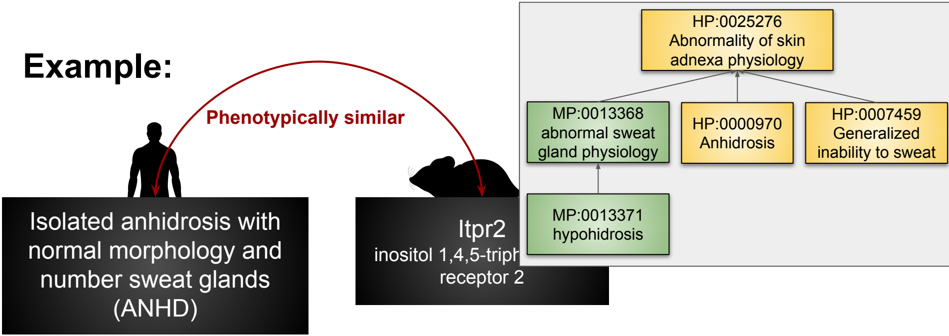
Anhidrosis

Mouse gene phenotypes:

abnormal sweat gland physiology

Hypohidrosis

Example:



Human disease phenotypes:

Generalized anhidrosis

Heat intolerance

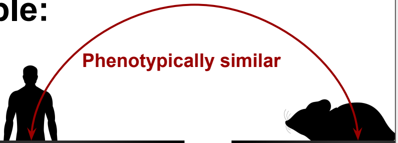
Anhidrosis

Mouse gene phenotypes:

abnormal sweat gland physiology

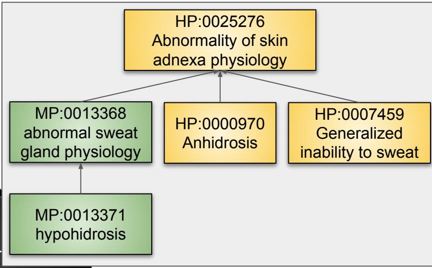
Hypohidrosis

Example:



Isolated anhidrosis with normal morphology and number sweat gla
(ANHD)

Itpr2
inositol 1,4,5-triph



How do we calculate the phenotypic similarity between genes and diseases using ontology embbeddings?

Human disease
phenotypes:

Generalized anhidrosis

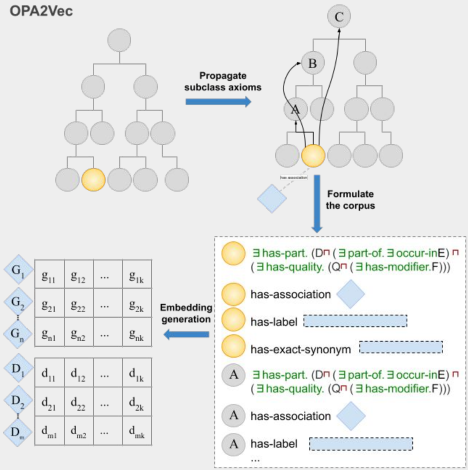
Heat intolerance

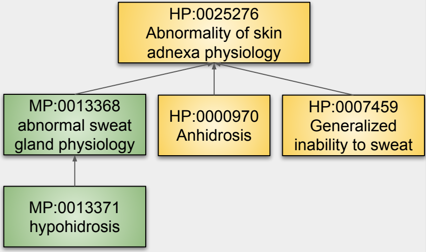
Anhidrosis

phenotypes:

abnormal sweat gland physiology

Hypohidrosis


OPA2Vec
Propagate subclass axioms
C
B
A
has association
Formulate the corpus
∃has-part. (D⊓ (∃part-of. ∃occur-inE) ⊓ (∃has-quality. (Q⊓ (∃has-modifier.F)))
has-association
has-label
has-exact-synonym
A ∃has-part. (D⊓ (∃part-of. ∃occur-inE) ⊓ (∃has-quality. (Q⊓ (∃has-modifier.F)))
A has-association
A has-label
...
Embedding generation
G1 g11 g12 ... g1k
G2 g21 g22 ... g2k
Gn gn1 gn2 ... gnk
D1 d11 d12 ... d1k
D2 d21 d22 ... d2k
Dm dm1 dm2 ... dmk

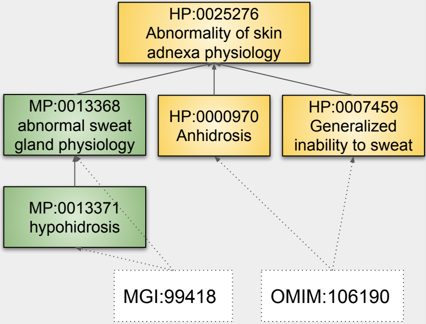

HP:0025276
Abnormality of skin adnexa physiology
MP:0013368
abnormal sweat gland physiology
HP:0000970
Anhidrosis
HP:0007459
Generalized inability to sweat
MP:0013371
hypohidrosis

OMIM:106190 has_annotaion **HP:0007459**
HP:0007459 has_label Generalized anhidrosis
HP:0007459 is_a **HP:0025276**
HP:0007459 has_synonyms Generalized anhydrosis
HP:0007459 has_synonyms Generalized inability to sweat

OMIM:106190 has_annotaion **HP:0000970**
HP:0000970 has_label Anhidrosis
HP:0000970 is_a **HP:0025276**
HP:0000970 has_database_cross_reference SNOMEDCT_US:39659002
HP:0000970 has_database_cross_reference UMLS:C0003028
HP:0000970 has_database_cross_reference MSH:D007007
HP:0000970 has_database_cross_reference SNOMEDCT_US:14662005
HP:0000970 has_database_cross_reference MEDDRA:10002512
HP:0000970 has_synonyms Anhydrosis
HP:0000970 has_synonyms Sweating dysfunction
HP:0000970 has_synonyms Sudomotor dysfunction,
HP:0000970 has_synonyms Lack of sweating

MGI:99418 has_annotaion **MP:0013368**
MP:0013368 Equivilant_to has_part some (functionality and (characteristic of some sweat gland) and (has modifier some abnormal))
MP:0013368 has_label abnormal sweat gland physiology
MP:0013368 has_definition Inability to sweat
MP:0013368 is_a **HP:0025276**
MP:0013368 has_synonyms sudomotor dysfunction
MP:0013368 has_synonyms abnormal sweat response
MP:0013368 has_synonyms sweating dysfunction

MGI:99418 has_annotaion **MP:0013371**
……



Word2Vec

Embedding

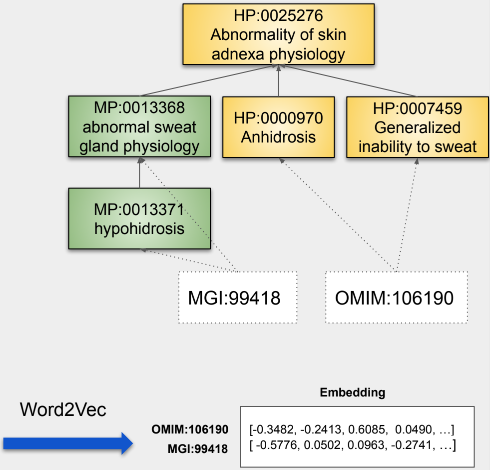
OMIM:106190	[-0.3482, -0.2413, 0.6085, 0.0490, …]
MGI:99418	[-0.5776, 0.0502, 0.0963, -0.2741, …]

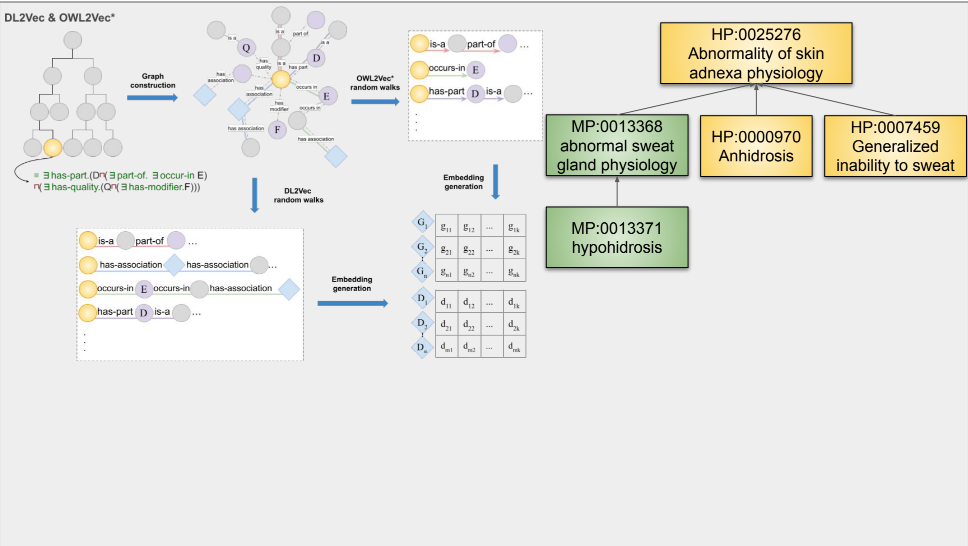
OMIM:106190 has_annotaion **HP:0007459**
HP:0007459 has_label Generalized anhidrosis
HP:0007459 is_a **HP:0025276**
HP:0007459 has_synonyms Generalized anhydrosis
HP:0007459 has_synonyms Generalized inability to sweat

OMIM:106190 has_annotaion **HP:0000970**
HP:0000970 has_label Anhidrosis
HP:0000970 is_a **HP:0025276**
HP:0000970 has_database_cross_reference SNOMEDCT_US:39659002
HP:0000970 has_database_cross_reference UMLS:C0003028
HP:0000970 has_database_cross_reference MSH:D007007
HP:0000970 has_database_cross_reference SNOMEDCT_US:14662005
HP:0000970 has_database_cross_reference MEDDRA:10002512
HP:0000970 has_synonyms Anhydrosis
HP:0000970 has_synonyms Sweating dysfunction
HP:0000970 has_synonyms Sudomotor dysfunction,
HP:0000970 has_synonyms Lack of sweating

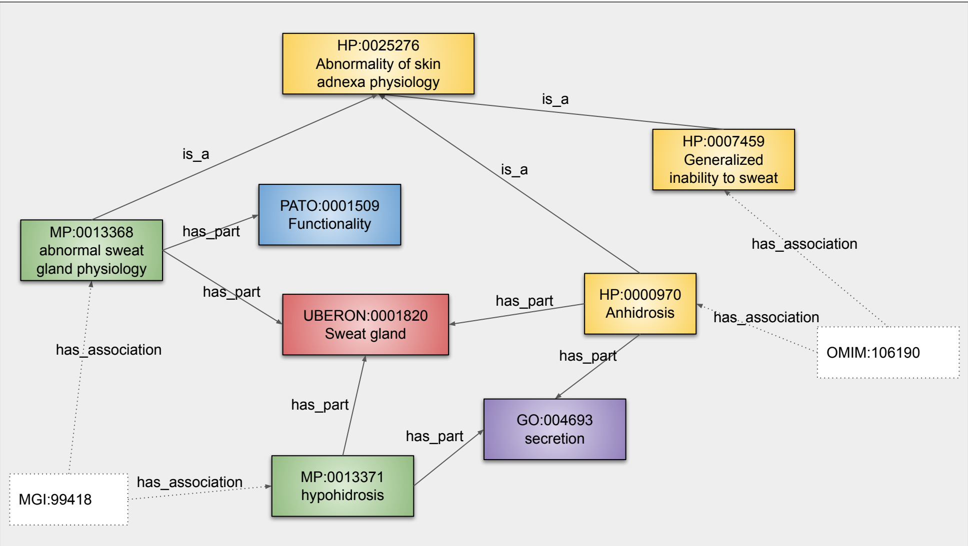
MGI:99418 has_annotaion **MP:0013368**
MP:0013368 Equivilant_to has_part some (functionality and (characteristic of some sweat gland) and (has modifier some abnormal))
MP:0013368 has_label abnormal sweat gland physiology
MP:0013368 has_definition Inability to sweat
MP:0013368 is_a **HP:0025276**
MP:0013368 has_synonyms sudomotor dysfunction
MP:0013368 has_synonyms abnormal sweat response
MP:0013368 has_synonyms sweating dysfunction

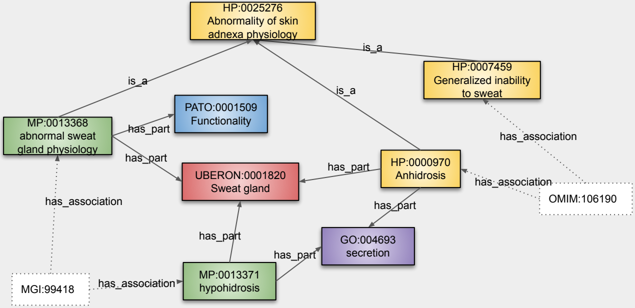
MGI:99418 has_annotaion **MP:0013371**
……

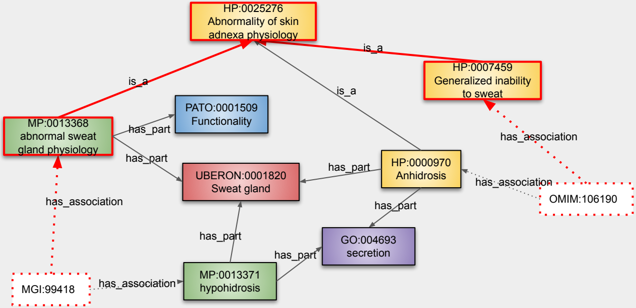




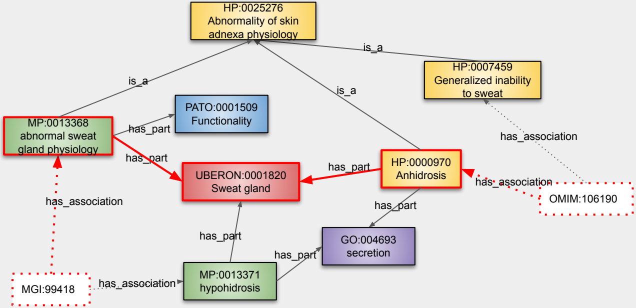
DL2Vec & OWL2Vec*
Graph construction
∃has-part.(D⊓(∃part-of. ∃occur-in E) ⊓(∃has-quality.(Q⊓(∃has-modifier.F)))
OWL2Vec* random walks
is-a part-of ...
occurs-in E
has-part D is-a ...
DL2Vec random walks
is-a part-of ...
has-association has-association ...
occurs-in E occurs-in has-association
has-part D is-a ...
Embedding generation
Embedding generation
HP:0025276 Abnormality of skin adnexa physiology
MP:0013368 abnormal sweat gland physiology
HP:0000970 Anhidrosis
HP:0007459 Generalized inability to sweat
MP:0013371 hypohidrosis


HP:0025276
Abnormality of skin adnexa physiology
is_a
HP:0007459
Generalized inability to sweat
is_a
is_a
PATO:0001509
Functionality
MP:0013368
abnormal sweat gland physiology
has_part
has_part
has_association
UBERON:0001820
Sweat gland
has_part
HP:0000970
Anhidrosis
has_association
has_association
OMIM:106190
has_part
has_association
has_part
GO:004693
secretion
has_part
MGI:99418
has_association
MP:0013371
hypohidrosis


HP:0025276
Abnormality of skin adnexa physiology
is_a
HP:0007459
Generalized inability to sweat
is_a
is_a
MP:0013368
abnormal sweat gland physiology
PATO:0001509
Functionality
has_part
has_part
has_association
UBERON:0001820
Sweat gland
has_part
HP:0000970
Anhidrosis
has_association
has_association
OMIM:106190
has_part
GO:004693
secretion
has_part
has_part
MGI:99418
has_association
MP:0013371
hypohidrosis

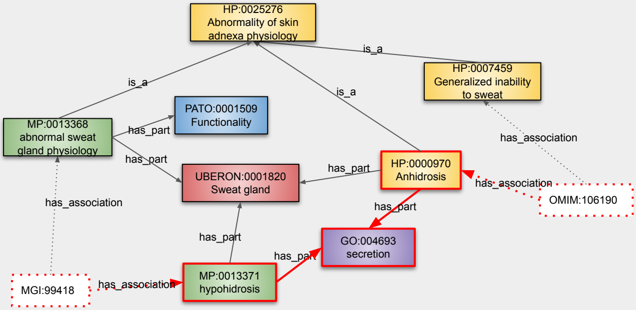


OMIM:106190 has_association HP:0007459 is_a HP:0025276 is_a MP:0013368 has_association **MGI:99418** …



OMIM:106190 has_association HP:0007459 is_a HP:0025276 is_a MP:0013368 has_association **MGI:99418** …

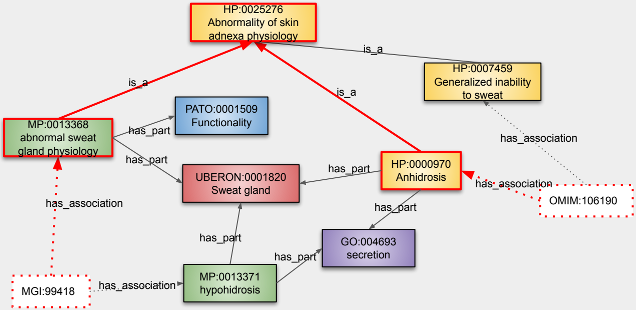
OMIM:106190 has_association HP:0000970 has_part UBERON:0001820 has_part MP:0013368 has_association **MGI:99418** …



OMIM:106190 has_association HP:0007459 is_a HP:0025276 is_a MP:0013368 has_association **MGI:99418** …

OMIM:106190 has_association HP:0000970 has_part UBERON:0001820 has_part MP:0013368 has_association **MGI:99418** …

OMIM:106190 has_association HP:0000970 has_part GO:004693 has_part MP:0013371 has_association **MGI:99418 …**

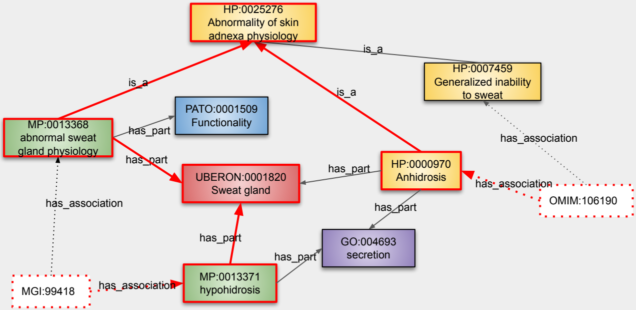


OMIM:106190 has_association HP:0007459 is_a HP:0025276 is_a MP:0013368 has_association **MGI:99418** …

OMIM:106190 has_association HP:0000970 has_part UBERON:0001820 has_part MP:0013368 has_association **MGI:99418** …

OMIM:106190 has_association HP:0000970 has_part GO:004693 has_part MP:0013371 has_association **MGI:99418 …**

OMIM:106190 has_association HP:0000970 is_a HP:0025276 is_a MP:0013368 has_association **MGI:99418** …



OMIM:106190 has_association HP:0007459 is_a HP:0025276 is_a MP:0013368 has_association **MGI:99418** …

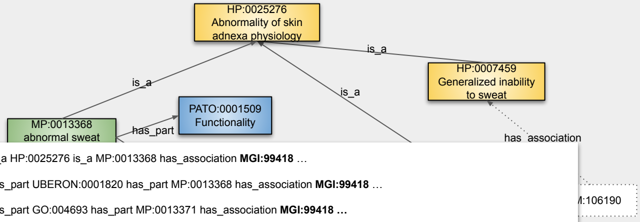
OMIM:106190 has_association HP:0000970 has_part UBERON:0001820 has_part MP:0013368 has_association **MGI:99418** …

OMIM:106190 has_association HP:0000970 has_part GO:004693 has_part MP:0013371 has_association **MGI:99418 …**

OMIM:106190 has_association HP:0000970 is_a HP:0025276 is_a MP:0013368 has_association **MGI:99418** …

OMIM:106190 has_association HP:0000970 is_a HP:0025276 is_a MP:0013368 has_part UBERON:0001820 has_part MP:0013371 has_association **MGI:99418** …

…



OMIM:106190 has_association HP:0007459 is_a HP:0025276 is_a MP:0013368 has_association **MGI:99418** …

OMIM:106190 has_association HP:0000970 has_part UBERON:0001820 has_part MP:0013368 has_association **MGI:99418** …

OMIM:106190 has_association HP:0000970 has_part GO:004693 has_part MP:0013371 has_association **MGI:99418 …**

OMIM:106190 has_association HP:0000970 is_a HP:0025276 is_a MP:0013368 has_association **MGI:99418** …

OMIM:106190 has_association HP:0000970 is_a HP:0025276 is_a MP:0013368 has_part UBERON:0001820 has_part MP:0013371 has_association **MGI:99418** …

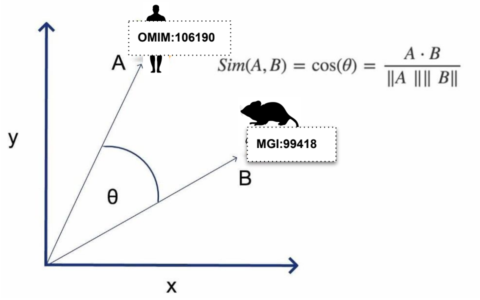
…

Embedding

OMIM:106190	[0.6909, 0.6992, 0.2646, -0.0663, …]
MGI:99418	[0.4071, 0.8932, -0.1988, -0.7038, …]

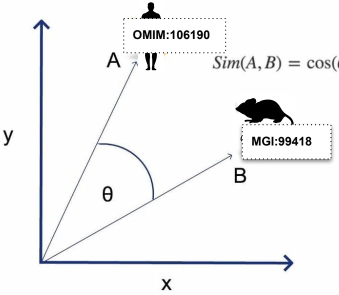
Calculating Phenotypic Similarity Approaches

- Unsupervised Approach
  - Cosine similarity



Calculating Phenotypic Similarity Approaches

- Unsupervised Approach
  - Cosine similarity



$$Sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

diseases

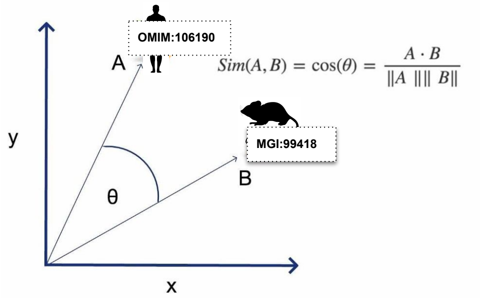
D_1	d_{11}	d_{12}	...	d_{1k}
D_2	d_{21}	d_{22}	...	d_{2k}
D_m	d_{m1}	d_{m2}	...	d_{mk}

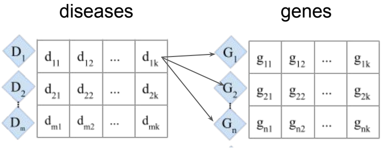
genes

G_1	g_{11}	g_{12}	...	g_{1k}
G_2	g_{21}	g_{22}	...	g_{2k}
G_n	g_{n1}	g_{n2}	...	g_{nk}

Calculating Phenotypic Similarity Approaches

- Unsupervised Approach
  - Cosine similarity

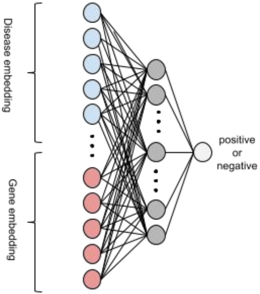




Predict associated genes to disease D_1:
$\max_{G \in genes}(Sim(D_1, G))$

Calculating Phenotypic Similarity Approaches

- Unsupervised Approach
  - Cosine similarity
- Supervised Approach
  - MLP



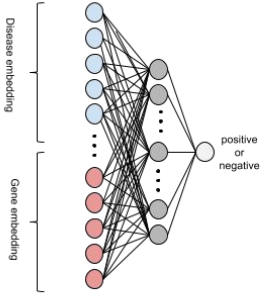
Calculating Phenotypic Similarity Approaches

- Unsupervised Approach
  - Cosine similarity

- Supervised Approach
  - MLP
    - Train/Test split

      10-fold cross validation

      Stratified by disease

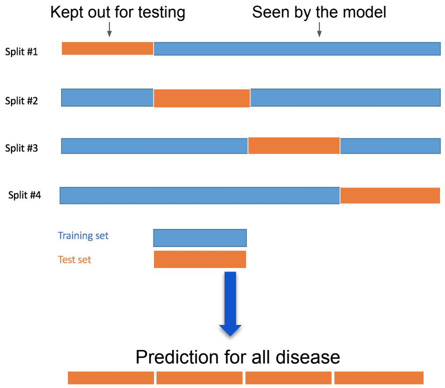


Calculating Phenotypic Similarity Approaches

- Unsupervised Approach
  - Cosine similarity
- Supervised Approach
  - MLP
    - Train/Test split

      10-fold cross validation

      Stratified by disease



Calculating Phenotypic Similarity Approaches

- Unsupervised Approach
  - Cosine similarity

- Supervised Approach
  - MLP
    - Train/Test split

      10-fold cross validation

      Stratified by disease

    - Positives/Negatives

Positives/Negatives imbalance …

Calculating Phenotypic Similarity Approaches

- Unsupervised Approach
  - Cosine similarity

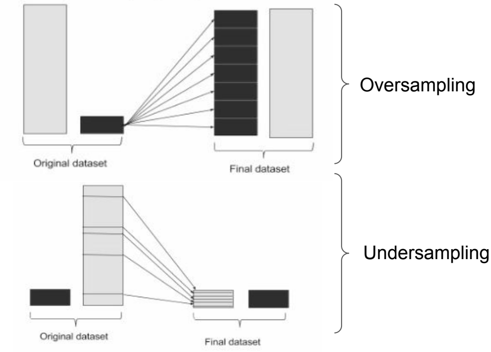
- Supervised Approach
  - MLP
    - Train/Test split

      10-fold cross validation

      Stratified by disease

    - Positives/Negatives

Positives/Negatives imbalance …



Calculating Phenotypic Similarity Approaches

- Unsupervised Approach
    - Cosine similarity

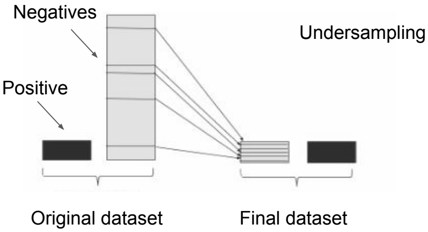
- Supervised Approach
    - MLP
        - Train/Test split

            10-fold cross validation

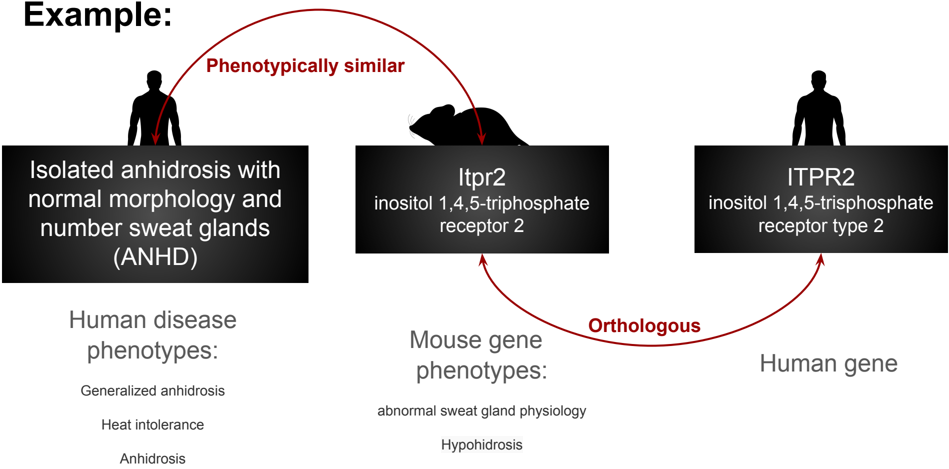
            Stratified by disease

        - Positives/Negatives

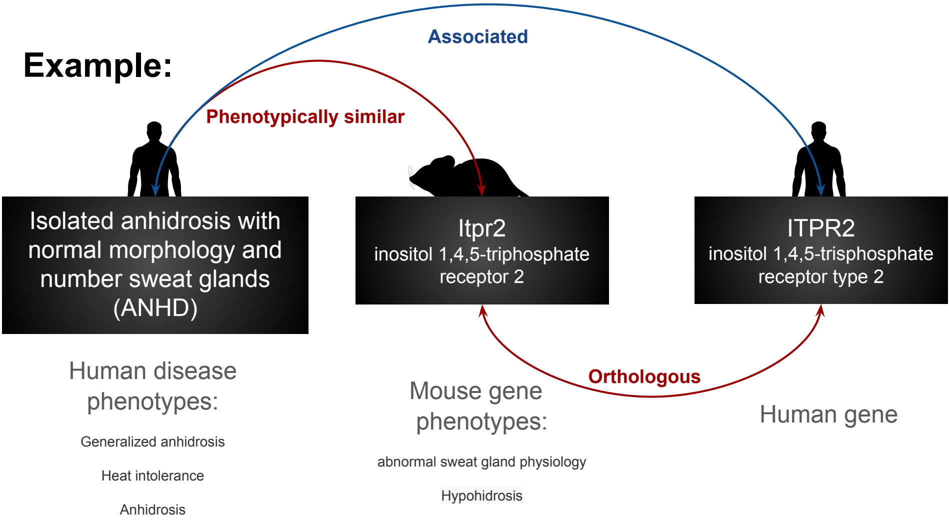
Positives/Negatives imbalance …



Example:


Phenotypically similar
Isolated anhidrosis with normal morphology and number sweat glands (ANHD)
Itpr2
inositol 1,4,5-triphosphate receptor 2
ITPR2
inositol 1,4,5-trisphosphate receptor type 2
Orthologous
Human disease phenotypes:
Generalized anhidrosis
Heat intolerance
Anhidrosis
Mouse gene phenotypes:
abnormal sweat gland physiology
Hypohidrosis
Human gene

Example:



Associated
Phenotypically similar
Isolated anhidrosis with normal morphology and number sweat glands (ANHD)
Itpr2
inositol 1,4,5-triphosphate receptor 2
ITPR2
inositol 1,4,5-trisphosphate receptor type 2
Orthologous
Human disease phenotypes:
Generalized anhidrosis
Heat intolerance
Anhidrosis
Mouse gene phenotypes:
abnormal sweat gland physiology
Hypohidrosis
Human gene

Hands On tutorial ..