DrugProt: Drug-Protein Relation Extraction Using Transformer-Based Models

Authors:
Volkan Bulca
İbrahim Özgürcan Öztaş

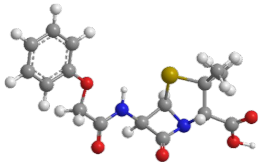
Advisor:
Assoc. Prof. Arzucan Özgür

Table of Contents

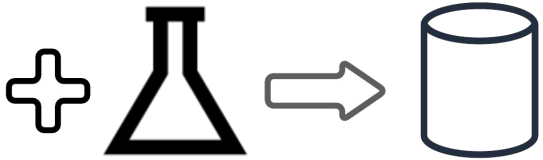
Motivation

Our motivation was to extract relations between drugs and proteins to

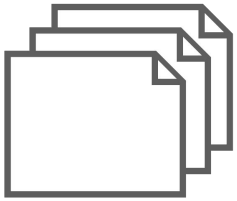
- Discover new drugs by using the existing ones
- Extract information about diseases
- Reorganize drugs for different purposes



Interactions between drugs and proteins

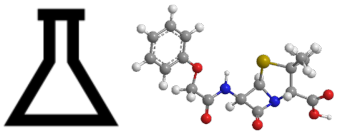


Search relation queries in DB



Formless biomedical texts

Manual Extraction



Information about drugs and proteins

Introduction

- We would like to express relations between drugs and proteins in an easily accessible, reshapable structures.
- To achieve such a task, we've *preprocessed and extracted the entity pairs* that have relations between its entities and *created a transformer-based model* to predict the future drug and protein pairs.

What is relation?

→ Raw Text: The h-OB were 10-100 fold more sensitive to DPHD than transformed osteoblasts: **DPHD** increased h-OB proliferation at 10nM and, at 100nM, activated **MAP kinase** signaling within 30min.

→ Entity Pair: DPHD - MAP kinase

→ Evidence: "... activated ..."

→ Relation: **ACTIVATOR**

→ Drug: **DPHD**

→ Protein: **MAP kinase**

What is relation?

→ Raw Text: During differentiation, **DPHD** promoted early expression of osteoblast transcription factors, RUNX2 and **osterix**.

→ Entity Pair: DPHD - osterix

→ Evidence: "... promoted ..."

→ Relation: INDIRECT-UPREGULATOR

→ Drug: **DPHD**

→ Protein: **osterix**

What is relation?

→ Raw Text: The reduction was more pronounced in the **fructose**-fed group and attributed to decreased hepatic expression of ACC2, FAS, SCD1, and **MTTP** and a decrease in the rate of hepatic triglyceride secretion.

→ Entity Pair: fructose - MTTP

→ Evidence: "... attributed to decreased …"

→ Relation: INDIRECT-DOWNREGULATOR

→ Drug: **fructose**

→ Protein: **MTTP**

What is relation?

→ Raw Text: The results showed that treatment with **EEDQ**, which blocked 80% to 85% of the **dopamine D2** and dopamine D1 receptors in substantia nigra, increased the levels of dopamine D2 receptor mRNA in substantia nigra by about 27%.

→ Entity Pair: EEDQ - dopamine

→ Evidence: "... blocked ..."

→ Relation: **INHIBITOR**

→ Drug: EEDQ

→ Protein: dopamine D2

Related Work

→ Traditional Stage(before 2010): Mostly SVMs and Random Forest

- Extracting protein - protein interaction information from biomedical text with SVM, 2006
- Exploiting shallow linguistic information for relation extraction from biomedical literature, 2006
- Evaluation of clustering algorithms for protein-protein interaction networks

→ DL Stage(2010-2019): CNNs, BiLSTMs, RNNs

- CNN-based Chemical-Protein Interactions Classification
- Identifying protein - protein interactions in biomedical literature using recurrent neural networks with long short-term memory, 2017
- Prediction of Protein-Protein Interactions with LSTM Deep Learning Model

→ Transformer Stage(2019-today): BERT, BioBERT, SciBERT

- BioBERT: a pre-trained biomedical language representation model for biomedical text mining

Overview

Objective: *Extract the relations between drugs and proteins* from the information provided in sentences.

Input Representation: *Sentence based approach* since nearly all sentences are within the model input range.

Multi-class Relation Extraction:

- Identify the relation type if *the sentence is stated as one that includes a relation.*

Model: BioBERT and Adam optimizer with lr=1e-5 & wd=3e-2

Dataset

→ DrugProt: It is released as part of a competition in BioCreative VII Track1.

→ Since the competition has just begun, we only have training dataset.

→ Abstracts file: PubMedId, Title, Abstracts

→ Entities file: PubMedId, RelativeEntityId, EntityType, StartingOffset, EndingOffset, EntityText

→ Relations file: PubMedId, RelationType, EntityArg1, EntityArg2

Abstracts

17512723 RDH12, a retinol dehydrogenase causing Leber's congenital amaurosis, is also involved in steroid metabolism. Three retinol dehydrogenases (RDHs) were tested for steroid converting abilities: human and murine RDH 12 and human RDH13. RDH12 is involved in retinal degeneration in Leber's congenital amaurosis (LCA). We show that murine Rdh12 and human RDH13 do not reveal activity towards the checked steroids, but that human type 12 RDH reduces dihydrotestosterone to androstanediol, and is thus also involved in steroid metabolism. Furthermore, we analyzed both expression and subcellular localization of these enzymes.

UTF-8 encoded TSV file with PubMed article abstracts.

- Article Identifier (PubMedId)
- Title of the Article
- Abstract of the Article

Entities

UTF-8 encoded TSV file with entities.

- Article Identifier (PubMedId)
- Entity Number
- Entity Type
- Starting Offset
- Ending Offset
- Entity Text

17512723	T1	CHEMICAL	466	480	androstanediol
17512723	T2	CHEMICAL	115	122	retinol
17512723	T3	CHEMICAL	9	16	retinol
17512723	T4	GENE-Y	219	230	human RDH13
17512723	T5	GENE-Y	232	237	RDH12
17512723	T6	GENE-Y	326	338	murine Rdh12
17512723	T7	GENE-Y	343	354	human RDH13
17512723	T8	GENE-N	139	143	RDHs
17512723	T9	GENE-Y	417	434	human type 12 RDH
17512723	T10	GENE-N	115	137	retinol dehydrogenases
17512723	T11	GENE-N	191	214	human and murine RDH 12
17512723	T12	GENE-Y	0	5	RDH12
17512723	T13	GENE-N	9	30	retinol dehydrogenase

Relations

UTF-8 encoded TSV file with relations between entities.

- Article Identifier (PubMedId)
- Relation Type
- Interactor argument 1
- Interactor argument 2

17512723 PRODUCT-OF Arg1:T1 Arg2:T9

Dataset

#abstracts:	3500
#entities:	89529
#relations:	17288

Relation Groups: Type by Type

- #INHIBITOR: 5392,
- #PART-OF: 886,
- #SUBSTRATE: 2003,
- #ACTIVATOR: 1429,
- #INDIRECT-DOWNREGULATOR: 1330,
- #ANTAGONIST: 972,
- #INDIRECT-UPREGULATOR: 1379,
- #AGONIST: 659,
- #DIRECT-REGULATOR: 2250,
- #PRODUCT-OF: 921,
- #AGONIST-ACTIVATOR: 29,
- #AGONIST-INHIBITOR: 13,
- #SUBSTRATE_PRODUCT-OF: 25

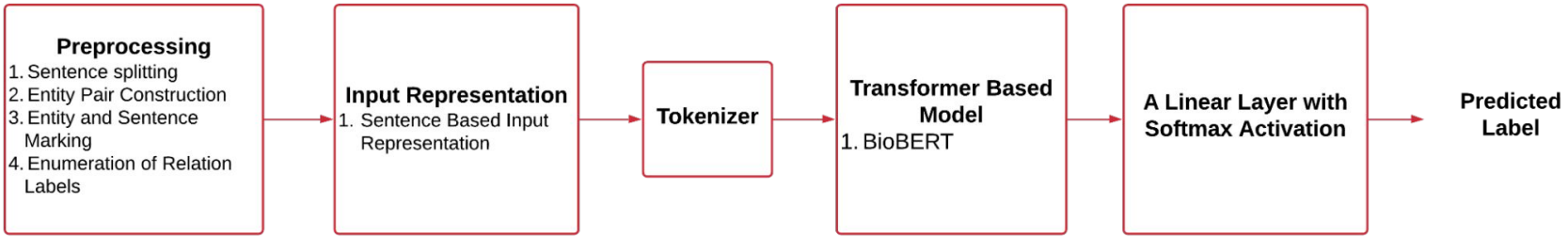
17512723 RDH12, a retinol dehydrogenase causing Leber's congenital amaurosis, is also involved in steroid metabolism. Three retinol dehydrogenases (RDHs) were tested for steroid converting abilities: human and murine RDH 12 and human RDH13. RDH12 is involved in retinal degeneration in Leber's congenital amaurosis (LCA). We show that murine Rdh12 and human RDH13 do not reveal activity towards the checked steroids, but that human type 12 RDH reduces dihydrotestosterone to androstanediol, and is thus also involved in steroid metabolism. Furthermore, we analyzed both expression and subcellular localization of these enzymes.

17512723	T1	CHEMICAL	466	480	androstanediol
17512723	T2	CHEMICAL	115	122	retinol
17512723	T3	CHEMICAL	9	16	retinol
17512723	T4	GENE-Y	219	230	human RDH13
17512723	T5	GENE-Y	232	237	RDH12
17512723	T6	GENE-Y	326	338	murine Rdh12
17512723	T7	GENE-Y	343	354	human RDH13
17512723	T8	GENE-N	139	143	RDHs
17512723	T9	GENE-Y	417	434	human type 12 RDH
17512723	T10	GENE-N	115	137	retinol dehydrogenases
17512723	T11	GENE-N	191	214	human and murine RDH 12
17512723	T12	GENE-Y	0	5	RDH12
17512723	T13	GENE-N	9	30	retinol dehydrogenase

... , but that human type 12 RDH reduces dihydrotestosterone to androstanediol, ...

17512723	PRODUCT-OF	Arg1:T1	Arg2:T9

Model Pipeline



Methods - Preprocess Stage

→ Sentence Splitting

- Used GENIA Sentence Splitter: Splitting the abstracts into its sentences in order.

→ Entity Pair Construction

- Check for each split sentence and find number of entity pairs in each sentence.
- Expand the sentences by that number and determine the locations of the entities.

Methods - Preprocess Stage

→ Entity and Sentence Marking

- Mark each Drug as **<e1>***Drug***</e1>** and each Protein as **<e2>***Protein***</e2>**.
- Mark start and end of sentences with special marks.
- Define special indexes for the marks within tokenizer: indexes {1,2,3,4,101,102} correspond to marks {<e1>,</e1>,<e2>,</e2>,“start”,“end”} respectively.

→ Enumeration of the Relation Labels

- Each label corresponding to 13 relations are enumerated with numbers {0,1,...,12}

Methods - Relation Extraction Stage

→ Tokenization of the Abstract Sentences

- Used the tokenizer of “dmis-lab/biobert-v1.1” and take only the abstracts having less tokens than 512.

→ Transformer Based Model

- Used the pre-trained BioBERT model in “dmis-lab/biobert-v1.1” from huggingface library.

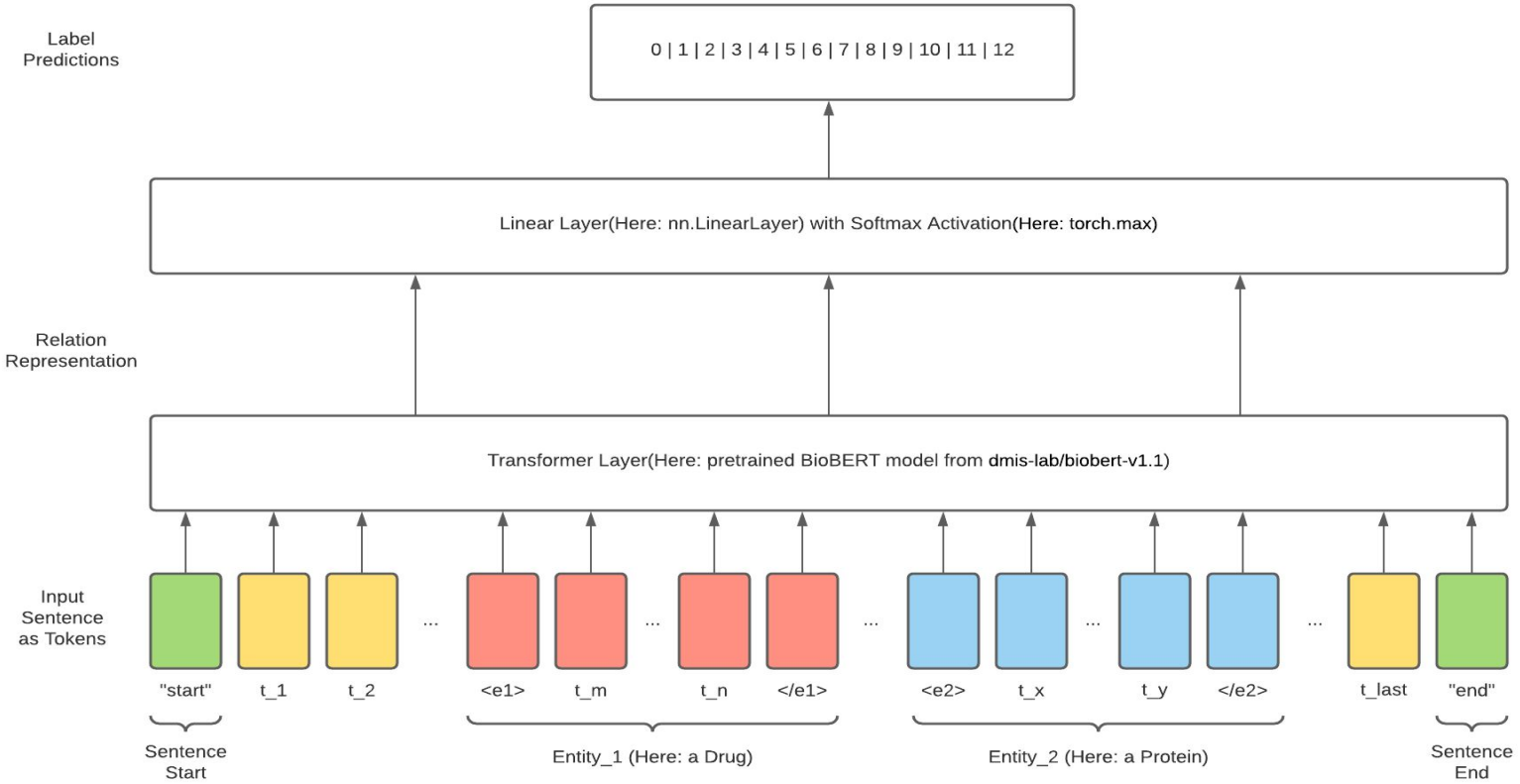
Methods - Relation Extraction Stage

→ Linear Layer and Softmax Activation

- Added a linear layer having 768 input features and 13 output features.
- Used softmax activation for most-likely relation extraction.

→ Prediction

- Predict the relation label of the entity pair-containing abstract according to the result of softmax activation and map it to corresponding relation name.


Label Predictions
0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12
Linear Layer(Here: nn.LinearLayer) with Softmax Activation(Here: torch.max)
Relation Representation
Transformer Layer(Here: pretrained BioBERT model from dmis-lab/biobert-v1.1)
Input Sentence as Tokens
"start"
t_1
t_2
...
<e1>
t_m
...
t_n
</e1>
...
<e2>
t_x
...
t_y
</e2>
...
t_last
"end"
Sentence Start
Entity_1 (Here: a Drug)
Entity_2 (Here: a Protein)
Sentence End

Results

→ Training Hyperparameters:

- Loss Function: Cross-Entropy Loss
- Learning rate: 1e-5
- Weight decay: 3e-2
- Optimizer: Adam

→ Evaluation Results on 0.2-ratio test set generated by given training dataset

- 3455 total predicted labels
- 0.77 overall accuracy
- 0.7633 weighted precision
- 0.7725 weighted recall
- 0.7660 weighted f1-score

Model evaluation results with lr=1e-5, wd=3e-2, 1 iteration only

	precision	recall	f1-score	# of true labels	# of predicted labels
'INHIBITOR'	0.86	0.84	0.85	1121	1094
'PART-OF'	0.58	0.80	0.67	148	206
'SUBSTRATE'	0.63	0.76	0.69	416	502
'ACTIVATOR'	0.68	0.70	0.69	244	254
'INDIRECT-DOWNREGULATOR'	0.67	0.73	0.70	266	289
'ANTAGONIST'	0.89	0.88	0.89	204	202
'INDIRECT-UPREGULATOR'	0.77	0.66	0.71	244	210
'AGONIST'	0.83	0.84	0.83	144	146
'DIRECT-REGULATOR'	0.84	0.72	0.77	504	433
'PRODUCT-OF'	0.66	0.54	0.60	146	119
'AGONIST-ACTIVATOR'	0.00	0.00	0.00	7	0
'AGONIST-INHIBITOR'	0.00	0.00	0.00	1	0
'SUBSTRATE_PRODUCT-OF'	0.00	0.00	0.00	10	0

Discussions & Conclusion

- Transformer based models are highly dependent on random token guessing, thus we should run our model 10 times with the same hyperparameters and average the results to lessen the effect of randomization.
- Using more generalized transformer based models would result in a lesser accuracy than our score due to the precision in the pre-training of the models. Generalized models are trained on vast data sets with different topics, thus its power to focus on one topic is highly decreased.
- Biomedical text mining highly depends on pre-trained models with high-precision fine tuning and our project shows one instance of such an occasion that BioBERT can be utilized and provides fairly strong results.

Future Work

→ Using different transformer-based pretrained models such as SciBERT, which is also a suitable for Relation Extraction tasks in scientific fields

→ Consulting different techniques for representing clustered relations such as graph networks

→ Training and testing our model with different hyperparameters to be able to eliminate the bias, deviations and randomness in the pre-trained model we used

→ We will continue working on this project for the BioCreative VII competition and seek to improve our results by trying the aforementioned suggestions of improvements.