Easy Semantification of Bioassays*

Marco Anteghini[1,2][0000−0003−2794−3853], Jennifer D'Souza[3][0000−0002−6616−9509], Vitor A.P. Martins dos Santos[1,2][0000−0002−2352−9017], and Sören Auer[3][0000−0002−0698−2864]

[1] Lifeglimmer GmbH, Markelstr. 38, 12163 Berlin, Germany
[2] Wageningen University & Research, Laboratory of Systems & Synthetic Biology, Stippeneng 4, 6708 WE, Wageningen, The Netherlands
{anteghini,vds}@lifeglimmer.com
[3] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
{jennifer.dsouza,auer}@tib.eu

A A Second Motivating Example for Bioassay Semantification

Assay ID 1061 As another example sentence, we consider 'The G-protein coupled formylpeptide receptor (FPR) was one of the originating members of the chemoattractant receptor superfamily. The present assay was undertaken to identify which of the 5 test compounds active in dose-response assays were FPRL1 antagonists.' While the sentence with various buried information units is not in a computable form, the logical statements 'has participant' → 'G protein coupled receptor' and 'has role' → 'target' about *G-protein coupled formylpeptide receptor* are as the semantic equivalent of information buried in the text. Generally, in semantified bioassays, the predicates 'has participant' and 'has role' occur frequently as a related logical statement sequences. The label 'has participant' often refers to a specific molecule participating in the bioassay, while the label 'has role' refers to the role of the participating molecule in the experimental process. In the case of the discussed example, it means the bioassay has a 'G protein coupled receptor' as a participant which is the *target* molecule in the experiment. Note, similar to our previous example, this information is not explicitly found in the text and would rely on a human annotator observing context and their background knowledge of the experiment. In this paper our aim is to expedite the annotation process with the help of machine learning over representative annotated examples.

B Comparison of Reported Results with our Prior Work

The dataset considered in this study has been created with a refined set of heuristics. Thus this dataset differs from the dataset in our earlier work. Note,

* Supported by TIB Leibniz Information Centre for Science and Technology, the EU H2020 ERC project ScienceGRaph (GA ID: 819536) and the ITN PERICO (GA ID: 812968).

our previous dataset had $|S| = 1756$ unique statements (after filtering for non-informative ones), however, this dataset has 1906 unique statements. In the following lines, we describe the differences in our datasets and thus explain the new results reported in this paper.

1. The ignored classes differ as shown in Table 1. In addition in the old dataset we removed three specific labels: 'has function $->$ aggregated', 'has participant $->$ Calcium', 'has participant $->$ 7-amino-4-methylcoumarin'.
2. In the dataset that we use in the present study, we have combined all occurrences of the 'has role' tag with its related tag pair. E.g. 'has inducer # has role $->$ Tetracycline # inducer', 'has participant # has role $->$ Propionaldehyde # substrate'.
3. previously we were including just the one with a BAO mapping ; here we have also other labels that are suitable for SciBert classification but are not in the BAO by checking the left hand manually and decided what to keep E.g. mapping: 'http://www.w3.org/1999/02/22-rdf-syntax-ns#type $->$ '10% fetal bovine serum', label: 'material entity culture serum' $->$ '10% fetal bovine serum'.

C Classification Bioassay Results

D Elbow Optimization for K

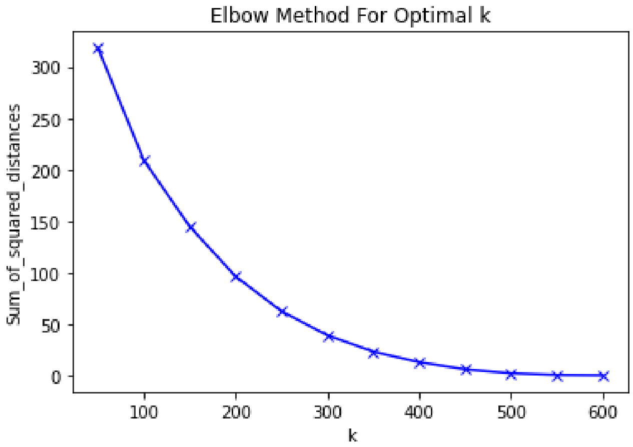

Fig. 1

Unique Statements LH	Old	New
‘has repetition point-number’	x	x
‘has concentration-point number’	x	x
‘has concentration value’	x	x
‘has endpoint’	x	x
‘has assay title’	x	x
‘has quality’	x	
‘has mode of action’	x	
‘has concentration unit’	x	
‘has response unit’	x	
‘has inducer’	x	
‘antibody’		x
‘has substrate’		x
‘screening campaign name’		x
‘has transcription factor’		x
‘material entity assay serum’		x
‘material entity culture medium’		x
‘material entity culture serum’		x
‘protein-protein’		x
‘screening campaign name’		x
‘Annotated by’		x
‘DMSO’		x
‘NCBI taxonomy ID’		x
‘PubChem TID’		x
‘absorbance wavelength’		x
‘cell modification temperature’		x
‘cell modification time’		x
‘construct DNA vector’		x
‘construct artificial regulatory region copy number’		x
‘construct gene ID’		x
‘construct organism’		x
‘enzyme reaction time’		x
‘gene ID’		x
‘gene mutation’		x
‘has assay medium’		x
‘absorbance wavelength’		x
‘has emission wavelength value’		x
‘has excitation wavelength value’		x
‘has incubation time value’		x
‘has signal direction’		x
‘has summary assay’		x
‘has temperature value’		x
‘is alternate confirmatory assay of’		x
‘is confirmatory assay of’		x
‘is counter assay of’		x
‘is identical assay of’		x
‘is lead optimization assay of’		x
‘is primary assay of’		x
‘is selectivity assay of’		x
‘positive control concentration’		x
‘substrate incubation temperature’		x
‘substrate incubation time’		x
‘uniprot ID’		x
‘material entity assay provider’		x

Table 1: Comparison among the old and the newly refined dataset. The Left-Hand parts of the labels are indicated.

false labels	*P*	*R*	*F*1
100	0.517	0.968	0.674
...	...	...	...
160	0.549	0.931	0.688
170	**0.600**	**0.939**	**0.729**
180	0.573	0.945	0.711
...	...	...	...
300	0.471	0.674	0.551

Table 2: Bioassay semantification results from five training optimization with different *false* classification instances (full table in appendix)

test set	*P*	*R*	*F*1
1st fold	0.600	0.939	0.729
2nd fold	0.573	0.956	0.713
3rd fold	0.589	0.936	0.719
Avg.	**0.588**	**0.944**	**0.720**

Table 3: Automatic bioassay semantification results from 3-fold cross validation with the optimal number of *false* classification labels (170).

	P	R	F1
top 10, 170RF	**0.53**	0.94	**0.67**
top 20, 170RF	0.50	0.89	0.64
top 30, 170RF	0.45	**0.95**	0.61
top 40, 170RF	0.37	0.94	0.52
top 50, 170RF	0.36	0.95	0.52
top 60, 170RF	0.41	0.92	0.57
top 70, 170RF	0.32	0.95	0.48
full dataset, 170RF	0.37	0.94	0.54
top 10, 180RF	**0.56**	0.93	**0.70**
top 20, 180RF	0.50	0.93	0.64
top 30, 180RF	0.49	0.94	0.65
top 40, 180RF	0.35	**0.96**	0.51
top 50, 180RF	0.39	0.94	0.55
top 60, 180RF	0.40	0.95	0.56
top 70, 180RF	0.37	0.95	0.53
full dataset, 180RF	0.35	0.94	0.51

Table 4: SCIBERT-based predictor results; 3-Fold CV on different subsets. The first column contains the number of Random False (RF) or *false* classification labels used in each bioassay for the analysed subsets (top 10, ..., full dataset) where each subset (e.g. top X) refers to the X most occurring Left-Hand (LH) part of a unique statement (e.g. top 10 contains the 10 most occurring LH of each unique statement). All the subsets where tested with 170 and 180 *false* classification labels.