THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

AUCKLAND BIOENGINEERING INSTITUTE

NLIMED: Natural Language Interface for Model Entity Discovery

Yuda Munarko, Dewan M. Sarwar, Anand Rampadarath, Koray Atalag, David Nickerson, The University of Auckland, New Zealand

Motivation

- Semantic annotation is used to ensure FAIR biosimulation models in biology and physiology.
- The COmputational Modeling in BIology Network (COMBINE) community recommends the use of the Resource Description Framework (RDF).
- The RDF provides the flexibility of model entity searching (e.g. flux of sodium across apical plasma membrane) by utilising SPARQL.
- Creating SPARQL is **not easy**.
- The availability of an interface to convert a natural language query to SPARQL is beneficial.

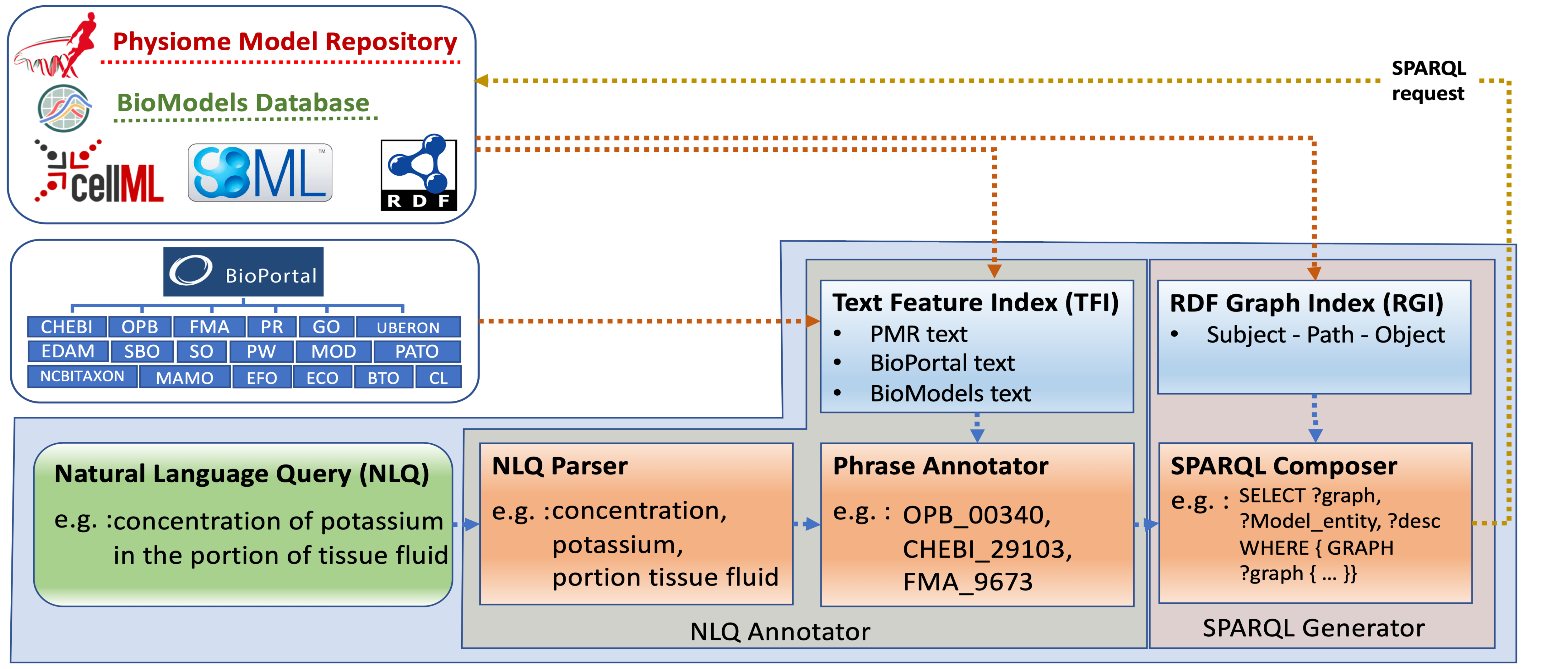
Results

- NLIMED, a natural language query to SPARQL interface to retrieve model entities from biosimulation models.
- The interface works for the PMR and BioModels.
- The interface has been implemented on the Epithelial Modeling Platform and Model Annotation and Discovery with the PMR.

Availability

- https://github.com/napakalas/NLIMED
- https://doi.org/10.1101/756304
- https://doi.org/10.17608/k6.auckland.11728977

Method


Physiome Model Repository
BioModels Database
cellML
SBML
RDF
SPARQL request
BioPortal
CHEBI
OPB
FMA
PR
GO
UBERON
EDAM
SBO
SO
PW
MOD
PATO
NCBITAXON
MAMO
EFO
ECO
BTO
CL
Text Feature Index (TFI)
• PMR text
• BioPortal text
• BioModels text
RDF Graph Index (RGI)
• Subject - Path - Object
Natural Language Query (NLQ)
e.g. :concentration of potassium in the portion of tissue fluid
NLQ Parser
e.g. :concentration, potassium, portion tissue fluid
Phrase Annotator
e.g. : OPB_00340, CHEBI_29103, FMA_9673
SPARQL Composer
e.g. : SELECT ?graph, ?Model_entity, ?desc WHERE { GRAPH ?graph { … }}
NLQ Annotator
SPARQL Generator

Current Implementation:

Epithelial Modelling Platform

Possible Users:

Figure1. NLIMED workflow. First, we create a Text Feature Index (TFI) and an RDF Graph Index (RGI) developed based on data on the PMR, BioModels, and BioPortal. Natural Language Query (NLQ), initially, is annotated into ontologies in Query Annotator module, then, translated into SPARQL in SPARQL Generator module.

NLQ Parser

- Convert a query to candidate phrases
- Utilising NLTK or Stanford parsers

Phrase Annotator

- Calculating the weight of candidate phrases to candidate ontologies W_{CO}
- Select the highest weight
- Avoid overlapping

$$W_{CO} = \sum_{i=term \, \epsilon \, phrase}^{n} \alpha \frac{p_i}{lp_i + nt} + \beta \frac{s_i}{ls_i + nt} + \gamma \frac{d_i}{(ld_i + nt)N} + \delta \frac{f_i}{(lf_i + nt)N} . log \frac{S}{S - ts_i}$$

Where:

- p_i, s_i, d_i = (1 or 0) preffLabel, synonym, definition
- lp_i, ls_i, ld_i, lf_i = the length of preffLabel, synonym, definition in class ontology and description in cellml or rdf.
- f_i, nt = the number of term in description, phrase
- N = the number of class ontologies having the term
- $ln \frac{S}{S - ts_i}$ = inverse document frequency,
- $\alpha, \beta, \gamma, \delta$ = multiple weighting scenario

RDF Graph Index

- Managing RDF in tree structures
- An index of Subject – Path – Object
- Path is a set of predicates

Text Feature Index

- Extracting features from repositories and BioPortal (preferred label, definition, synonym, description)
- Implement inverted index for fast retrieval

SPARQL Composer

- Constructing SPARQL based on Text Feature Index and Phrase Annotator results.

Summary

NLQ Annotator can identify class ontologies in NLQ

Method	Precision	Recall	F-measure	Query accuracy	Exec time
NLQ Annotator + Stanford parser	0.744	0.768	0.756	0.549	0.532
NLQ Annotator + NLTK parser	0.591	0.728	0.652	0.333	0.101
NCBO Annotator	0.402	0.376	0.388	0.196	36.697

NLIMED can handle a wide range of NLQ types containing one or many terms with one or many phrases.

SPARQL Generator storing RDF graph as indexes can generate all possible SPARQL based on provided ontologies.

Future Works

Implement NLIMED on:

- The Physiome Model Repository (PMR) aggregated search feature
- A generic biosimulation model search engine accommodating PMR and BioModels

Further, we are interested to explore lexical semantics inside NLQ and semantic concepts inside model entities to increase NLIMED performance and its use as a question and answer system.

References

Neal, M.L., König, M., Nickerson, D., Mısırlı, G., Kalbasi, R., Dräger, A., Atalag, K., Chelliah, V., Cooling, M.T., Cook, D.L. and Crook, S., 2019. Harmonizing semantic annotations for computational models in biology. *Briefings in bioinformatics*, *20*(2), pp.540-550.