A Pipeline for Analysis of Published Abstracts for Information on Protein-Protein Interactions

Maurice H.T. Ling, Christophe Lefèvre, Kevin R. Nicholas

CRC for Innovative Dairy Products, Department of Zoology, University of Melbourne, Parkville, 3010 Victoria

Abstract: There have been large volumes of literature published on mouse intracellular protein-protein interactions. However, there have been little attempts to aggregate these information into a usable model or concept map of protein-protein interactions, such as protein-protein bindings and activations. We have established a process for the handling and analysis of published abstracts from PubMed to extract information on protein-protein interactions, using only open source software and tools. A Firebird database, Muscopedia, which forms the central point of this pipeline, is used to store the abstracts and its processed forms. Muscopedia is interfaced to Python programming language through a Python DB-API compliant library, kinterbasdb. Abstracts were downloaded from PubMed using NCBI's Simple Object Access Protocol (SOAP) server and scanned for abbreviations using BioNLP server in Stanford University. Terms in the abstracts were substituted for their abbrevations , for example, the term "SOCS" is substituted for "suppressor of cytokine signaling", before text processing using MontyLingua. MontyLingua is a natural language processing kit, which has 'commonsense' built in and is used to process each abstract into a list of subject-verb-object (SVO) structure. On average, each abstract will be processed into 30 to 40 SVOs. Information on protein-protein interactions can then be extracted from this set of SVOs by using suitable verbs. These information will then be used to construct a concept map of protein-protein interactions.

Flow Diagram Illustrating the Main Processing Steps
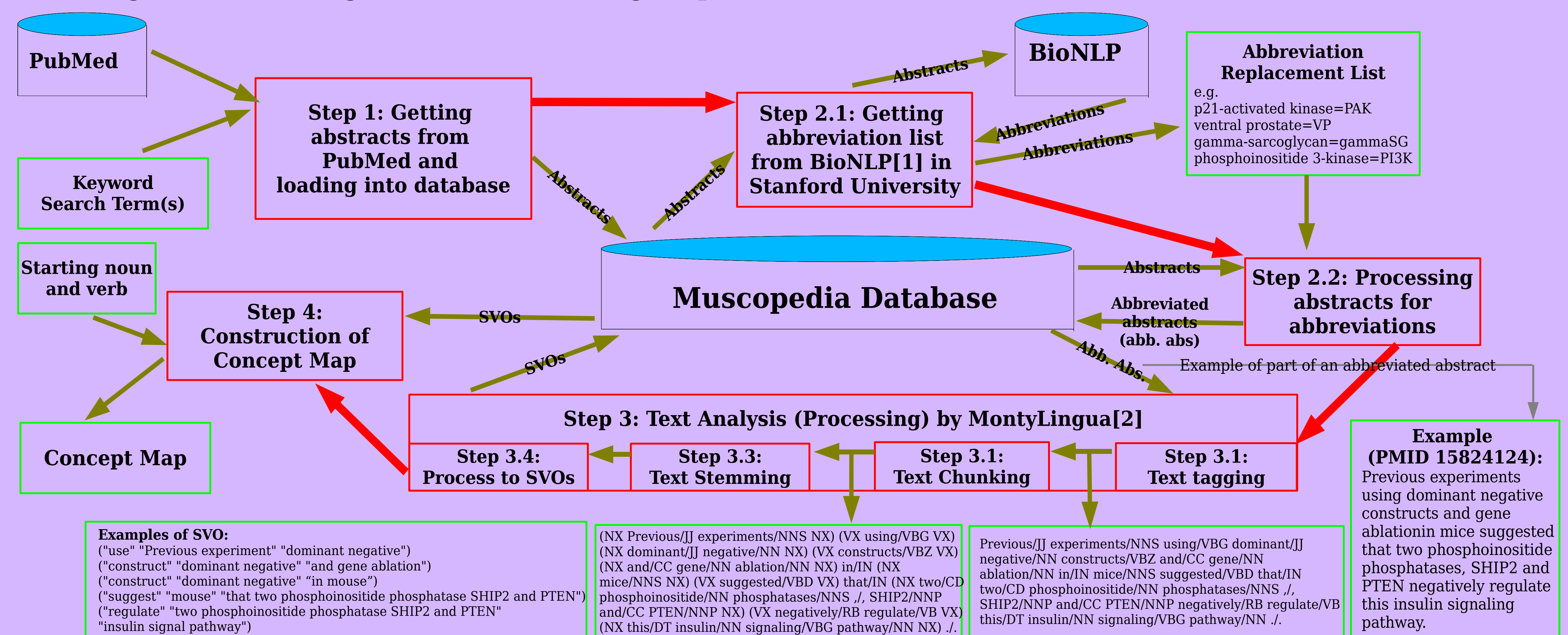
Result Preliminary Activation and Binding Pathways of Insulin and Related Pathways Constructed from Analyzed Text (Green lines represents activation reactions and blue lines represents binding reactions)

Conclusions
1. The pipeline had demonstrated core capabilities in managing and processing abstracts for the construction of concept maps.
2. Using insulin activation and binding, the ability of constructing a concept map from published abstracts has been demonstrated.

Future Work
1. Currently, the construction of concept map is largely a manual process. Possibilities of user-directed, semi-automation of construction of concept maps will be explored.
2. The current pipeline uses only one text analysis engine (MontyLingua). Improvements may be made possible by synergistic use of other text mining and analysis engines.

References
[1] Chang JT, Schutze H, and Altman RB. Creating an online dictionary of abbreviations from MEDLINE. J Am Med Inf Assoc 9: 612-620, 2002.
[2] Liu H. MontyLingua: An end-to-end natural language processor with common sense, 2004. (web.media.mit.edu/~hugo/montylingua)