

Extracting Chemical CYP proteins interactions from Literature Using Natural Language Processing Methods

D Jiao¹

D Wild²

¹ *School of Informatics, Indiana University at Bloomington, Bloomington USA*

² *School of Informatics, Indiana University at Bloomington, Bloomington USA*

This poster describes the development of an information extraction system which maps interactions between chemicals and CYP proteins from existing literature, using machine learning and natural language processing methods. The interaction between CYP proteins and chemicals is important in drug discovery and development. In this system, abstracts from articles related to CYP and chemical interactions are preprocessed using named entity recognition methods to identify chemical names and CYP names, with the help of dictionaries generated from biological and chemical ontologies. Chemical structures are also attached to chemical names for future processing. The texts are then parsed by a syntactic parser to create a dependency graph in which grammatical relationships between constituents of the sentences are generated. Then interactions between CYP and chemicals are extracted by identifying certain keywords, together with the protein and chemical names based on the dependency graph. The extracted information, including the chemical compounds, their structures, the proteins, and the interactions between chemicals and proteins are stored in a database for retrieval and further analysis. In this poster, the training process to build certain components of the system, problems encountered during the system creation, and the evaluation of the system will be discussed in detail.

References:

1. Corbett, P.; Murray-Rust, P. High-Throughput Identification of Chemistry in Life Science Texts. *Computational Life Sciences II*. **2006**, 107-118.
2. Clegg, A. B.; Shepherd, A. J. Benchmarking natural-language parsers for biological applications using dependency graphs. *BMC Bioinformatics* **2007** 8, 24+.
3. Feng, C.; Yamashita, F.; Hashida, M. Automated Extraction of Information from the Literature on Chemical-CYP3A4 Interactions. *J Chem Inf Model* **2007** 47 (6), 2449 -2455.
4. Kulick, S.; Bies, A.; Liberman, M.; Mandel, M. ; McDonald, R.; Palmer, M.; Schein, A.; Ungar, L.. *Integrated Annotation for Biomedical Information Extraction*. HLT/NAACL, 2004.