Automatic Generation of BEL Statements from Text-mined Biological Events

Haibin Liu[1§], William A Baumgartner Jr[2§], Natalie Catlett[3], Andrea Matthews[3], Phoebe Roberts[4], Christophe Roeder[2], Aaron Van Hooser[3], Daniel Ziemek[4], Lawrence E. Hunter[2], K. Bretonnel Cohen[2]

[1] NCBI, Bethesda, MD, USA
[2] University of Colorado School of Medicine, Aurora, CO, USA
[3] Selventa, One Alewife Center, Cambridge, MA 02140, USA
[4] Computational Sciences Center of Emphasis, Pfizer Worldwide Research & Development, Cambridge, MA, USA

[§]Both authors contributed equally to this work.

The Biological Expression Language (BEL; http://www.openbel.org/) designed by Selventa^{TM} aims to represent scientific findings in the life sciences with a focus on capturing causal relationships. Knowledge in BEL is represented as BEL statements that form the basis for several successful approaches to interpreting transcriptional and genetics data in light of prior causal knowledge. However, the manual curation of BEL statements remains a bottleneck of the subsequent reasoning process. Text mining has made significant progress in recent years in extracting semantic events involving genes and proteins, such as binding and regulatory events, from the biomedical literature. Since nested event structures correspond to the causal relationship chains in BEL, automatic generation of BEL statements from text-mined biological events is feasible.

This work reports our initial attempt to bridge biological events with causal BEL statements. A knowledge-driven, surjective mapping schema is proposed to transform events of nine GENIA event types into BEL statements. A probability-based confidence score is assigned to each statement. The Turku event extraction system is used to extract biological events from Medline abstracts. Protein mentions are subsequently normalized to concepts in the Protein Ontology. When applied to the EVEX database consisting of events extracted by the Turku system over the 2009 distribution of Medline, our conversion results in a large-scale set of 126,880 BEL statements in 90,620 sentences from 63,434 Medline abstracts. In addition, evaluation of our protein normalization system against the CRAFT corpus shows a 0.81 precision and a 0.47 recall when using an overlapping-span matching criterion.