Probabilistic Annotation Framework: Knowledge Assembly at Scale with Semantic and Probabilistic Techniques

Ross King[1], **Szymon Klarman**[2], **Larisa Soldatova**[2] and **Robert Stevens**[1]

{ross.king | robert.stevens}@manchester.ac.uk
{szymon.klarman | larisa.soldatova}@brunel.ac.uk

[1]School of Computer Science, University of Manchester
[2]Department of Computer Science, Brunel University London

We report on the ongoing efforts in conducting large-scale knowledge assembly in the context of the *Big Mechanism* research program. This aims to develop an artificial intelligence infrastructure for automatic construction of large, mechanistic knowledge models in the cancer biology domain, following the *read – assembly – explain* pipeline: 1) extracting facts about relevant molecular interactions from thousands of PubMed publications using text-mining technology; 2) assembling these facts into large-scale models, further enriched with data from existing structured sources; 3) explaining the findings about the underlying mechanisms and suggesting interesting, experimentally testable hypotheses.

The major challenge of the assembly phase, addressed in this work, lies in the prevalence of uncertainty concerning the collected information, which is caused by a range of factors, such as ambiguity of entities and relationships detected; inaccuracy of extraction; trust and provenance; etc. The key principle underlying our proposed *Probabilistic Annotation Framework*, being developed to support the knowledge assembly process, is to represent all such uncertainty-related information explicitly alongside the extracted facts qualified by it. Technically, the framework is based on the intertwined applications of semantic and probabilistic techniques. In particular, we have developed dedicated OWL ontologies for describing biological entities and their interactions, as found in the text, together with their associated provenance and probabilistic annotations. We then reason over this knowledge using probabilistic logic programs (supported by the ProbLog system) in order to determine the likelihood of particular findings and to perform probabilistic updates on the model. The probabilities are thus a first-class citizen in the model. In this way the resulting knowledge base can be constantly revised or updated on the arrival of new evidence, as well as providing meaningful explanations as to why and to what extent it endorses the truth of the contained statements.