

ABSTRACT

GlyGen is an international glycoinformatics project developed to facilitate glycobiology research through genomic, protein and glycan specific data integration and data management. The major resources for site-specific protein-glycan data in GlyGen are UniCarbKB, GlyTouCan, and UniProtKB. Currently, the amount of glycan-protein data available from these databases or other similar resources is limited. With the new technological advancements in the field of glycobiology, there is a huge influx of data being made available through publications that need to be captured and integrated with relevant molecular biology data. One approach is to use text-mining to capture relevant information and to provide additional annotations to the existing data. Leveraging our previous work in developing similar tools such as RLIMS-P (Rule-based Literature Mining System for protein Phosphorylation), we propose to develop a text-mining tool to capture glycan-protein and contextual data from new and existing publications. We intend to capture important information specific to glycosylation by focusing on all aspects such as glycoprotein (e.g. expression system, source, and species), mass spectrometry information (e.g. analyzer, proteolytic enzymes), glycan (e.g. structure, composition) and residue information (e.g. protein site, amino acid). This data will be used to not only extend the amount of knowledge in GlyGen but also help in improving existing annotations retrieved from collaborating resources mentioned above. The data gathered through this project will be made available to users using appropriate evidence tags from GlyGen portal. Additionally, the packaged software in a Docker Container will be made available to the GlyGen team for reuse and dissemination. This project has the potential to significantly expand the data available through GlyGen, increasing the usefulness of the system, and change the future course of data extraction and annotation for GlyGen.