

TITLE

Develop automatic literature mining tool for extracting context specific glycan-protein information that will enhance the extent and quality of data in GlyGen.

AIMS

Glycosylation is one of the major post-translational modifications as it affects several physical and chemical properties of a protein, including protein folding, stabilization and protein interactions. Given the complex nature of this process, the interpretation of the experimental data into relevant glycobiology knowledge has been slow and challenging. However, over the last several years, rapid technological advancements and growing interest in understanding glycan-protein-gene relationships, has led to a significant increase in our knowledge of this phenomenon and an increasing number of publications describing it. Using our experience in developing text-mining algorithms, we propose to develop a tool to capture site-specific glycosylation and other relevant glycoprotein data from publications.

Following are the major aims.

Aim 1: Develop and run text-mining software that can extract information from publications which can be used to improve GlyGen content.

- a) Develop and use automatic text-mining tool on the existing publications in GlyGen (sourced from UniProtKB, UniCarbKB, GlyTouCan) as a method to add evidence tags to existing annotations as well as capture additional information.
- b) Increase site-specific and protein specific glycan data in GlyGen by adding new glycosylation sites and additional annotations such as glycosylation type, glycan structures, source tissue/cell line, and disease through literature mining of new publications (publications not yet in GlyGen).

Aim 2: Replicability and usability of the tool.

- a) Generate a list of glycan specific terms from publications that can be subsequently leveraged to identify the structural features or compositions of the reported glycans.
- b) Provide text-mining software as a deliverable so that it can be run automatically on new publications.

The extracted data, as well as data generated through inference, will be represented as GlyGen data objects and will be accessible through GlyGen Data page (<http://data.glygen.org/browse>). The data will also be accessible through GlyGen protein and glycan search pages (<http://glygen.org>) where users can perform additional queries, searches, as well as view and download the data. This project is expected to significantly expand GlyGen data, its quality and usefulness, and transform the future course of data collection and annotation in GlyGen.