Data and Services Discovery projects - Transformative Data Collections

Title

Improvements in standardisation of storage and representation of glycomics data collections

Approach

During this funding period we focused on expanding the coverage and diversity of annotated glycoproteins stored UniCarbKB (http://www.unicarbkb.org) and improved the use of data standards and common ontologies to describe data collections. Our approach adds value to research data by providing high quality metadata that describes the provenance, acquisition and interpretation of data using best practices recommended by MIRAGE. To describe the data collections, we adopted and extended ontologies developed by the glycoscience community including GlycoRDF and GlyCoCoO. Data discoverability has been improved by embedding structured data markups into the data landing pages. Finally, the legacy web service API will be replaced with two new querying resources: i) a semantic SPARQL interface, and ii) a flexible GraphQL engine.

This project involved participants from the Institute for Gycomics (Griffith University), GlyGen (George Washington University, USA), and BTI A*STAR Singapore. The major project outcomes are:

Presentations
The UniCarbKB-GlyGen component of the work funded by ARDC was presented at the 2nd Australasian Glycomics Symposium (HUPO pre-conference satellite event, Adelaide 14th September; https://www.hupo2019.org/2nd-ags/), and the annual NIH GlyGen meeting (Washington DC, 21st-23rd August).

Data Collections
The glycoproteomics data collections described in the accompanying FAIR spreadsheet have been included in the GlyGen data feed, which will be validated and released to the public in the next major release (November 2019). These data collections are also available in the UniCarbKB triplestore based on the Glycoconjugate Ontology (GlyCoCoO) which has been designed to standardise the annotation of glycoconjugate data.

As part of an on-going collaboration with A*STAR Singapore we have added a GSL data collection to the GlycoStore database (https://www.glycostore.org/collections) which (for the first time) provides access to ion mobility and separation data.

Web Applications and Services

GlyCoCoO is an international effort serving as a basis for data exchange formats, the ARDC support has allowed us to complete the first major release with a publication in preparation. The SPARQL endpoint (http://sparql.unicarbkb.org) is hosted on ARDC infrastructure and accompanying documentation provided at https://app.gitbook.com/@unicarbkb-glycostore/s/data/. Full details and access to the triplestore data files can be accessed from https://github.com/glycoinfo/GlycoCoO.

The funding received led to the implementation of a GraphQL service, which has been designed to improve data access to UniCarbKB and GlycoStore for developers. This resource is hosted on ARDC infrastructure and a publication is in preparation, access and documentation will be provided upon submission.

Additional outputs include a new curated cell line platform (published and unpublished data collections) which is available for review by ARDC.

FAIR

Refer to FAIR assessment spreadsheet.

Collaboration and coverage

As described above this project is highly collaborative and brings together researchers from leading national and international research institutions. Due to the short period consumer analytics have not be acquired e.g. visitor numbers, however, by improving the content/layout of data landing pages the deposition of data has increased the visibility of the data collections. The project and presentations at international conferences has strengthened our collaborations with international glycoinformatic and analytical researchers, and has helped commence discussions with national partners and activities to establish an Australian Platform for Glycoscience.

Sustainability

UniCarbKB and its affiliated resources are key partners of the NIH GlyGen project (https://www.glygen.org) and the community driven initiatives GlycoRDF and GLIC. As part of the GlyGen project all past and future curated UniCarbKB and GlycoStore data collections are shared, and a data pipeline between the institutions is available including: a QC/QA process and documented API's for retrieve data stored. In the long-term core attributes of data collections will be accessible from other resources including UniProt and NCBI.

To support service sustainability, we provide Docker images for UniCarbKB and GlycoStore, and all code is publicly available. Beyond this funding period we aim to continue promoting the work described and maintaining access to the data collections. In conjugation with the MIRAGE initiative our goal is maintain high quality and well-described data collections that follow FAIR

principles providing users with specific, technical descriptions of the metadata required to evaluate the analysis, and thus as a basis for making judgments regarding the validity of specific conclusions in the manuscript. To this end, data collections will be freely accessible beyond this project and links retained with existing partnering institutions and additional funding will be sought.

Learnings

A key lesson learned is the need for controlled vocabularies/ontologies to describe data collections, but the sheer volume of ontologies available poses problems when making a correct decision, especially in the long-term. A major hurdle within the glycosciences has been the inconsistencies in reporting structural data and the diversity of experimental methodologies. This poses challenges when bringing together data and subsequently performing queries. Recently, Wu and colleagues (Future Directions in Data Discovery, eResearch Australasia 2019) stated the benefits of structured metadata in landing pages to enhance data discovery. This offers an approach for federated search across repositories - it would advantageous for the ARDC to explore this (bio)schema.org approach at a national level and consider best practices, which may lead to growth of transformative collections.

Impact

In this short period of funding five publications are in preparation or under review. The manuscripts describe: i) SPARQL endpoints and the GlycoCoO ontology for standardising glycoprotein data collections (led by Soka University, Japan); ii) a GraphQL platform; iii) biocuration processes and adoption of ontologies; iv) a database of glycan profiled cancer cell lines; and v) a software application for capillary electophoresis analysis interfaced with GlycoStore. Finally, a review of software and databases available in the glycomics space has been submitted.

This funding has enabled us to strengthen our collaborations with international glycoinformatic and analytical communities, and also helped us commence discussions with national partners. Importantly, a new research project to improve content and access of cancer cell line glycome data is under way, and a glycan-array database for storing host-pathogen interactions is being pursued.

The long-term benefits of this project will add value to the existing national investment in glycoscience research infrastructure. In partnership with international partners the project is committed to improving access and enabling research workflows on a national and international scale by maintaining accessibility to an eResearch infrastructure. It offers excellent opportunities to make use of exploitable results, both in the context of this project and other programs. The project will directly impact the implementation of the first centralised resource with the capacity and capability organise, integrate large sets of glycomics/glycoproteomics data impacting the ability of researchers to extract meaningful information, and, therefore, identify relationships between disease related proteins and glycosylation. By providing a sustainable eResearch infrastructure, in the long-term, it will significantly enhance and benefit a diverse range of on-going life science research activities – it will be possible to ask broader biological questions that

straddle glycomics, proteomics and genomics. The return on this investment will include more productive ways to create new informatic and data science technologies, as well as complementing analytical techniques for tacking societal challenges focused on Health, demographic change and wellbeing.

Report prepared by: Matthew Campbell, Institute for Glycomics, Griffith University
Date: 16th October 2019