

Glycans, the forgotten biomolecular actors of the big picture

Matthew P. Campbell¹✉, Julien A. Mariethoz², Catherine M. Hayes³, Pauline G. Rudd⁴, Niclas G. Karlsson³, Nicki H. Packer¹, Frédérique Lisacek²

¹Biomolecular Frontiers Research Centre, Macquarie University, Sydney, Australia

²Proteome Informatics Group, Swiss Institute of Bioinformatics, Geneva, Switzerland

³Department of Biomedicine, Gothenburg University, Gothenburg, Sweden

⁴NIBRT, Dublin, Ireland

Motivation and Objectives

Glycans or carbohydrates, both in the form of polysaccharides or glycoconjugates are increasingly recognised as being implicated in human health. Glycosylation is probably the most important post-translational modification in terms of the number of proteins modified and the diversity generated. Since glycoproteins, glycolipids and glycan-binding proteins are frequently located on the cell's primary interface with the external environment many biologically significant events can be attributed to glycan recognition. In other words, glycans mediate many protein-protein interactions. In spite of such a central role in biological processes, the study of glycans remains isolated, protein-carbohydrate interactions are rarely reported in bioinformatics databases and glycomics is lagging behind other -omics.

Recent progress in method development for characterising the branching structures of complex carbohydrates has now enabled high throughput technology. Automation then calls for software development. Adding meaning to large data collections requires bioinformatics means. Current glycobioinformatics resources do cover information on the structure and function of glycans, their association with proteins or their enzymatic generation. However, this information is partial, scattered and often inaccessible to non-glycobiologists.

In partnership with expert international research groups we are involved with the development of the UniCarb KnowledgeBase (UniCarbKB), an effort to develop and provide an informatic framework for the storage and the analysis of high-quality data collections on glycoconjugates, including informative metadata and annotated experimental datasets (Campbell et al., 2011). UniCarbKB is an initiative designed to support research in systems biology by complementing proteomics with glycomics

Methods

To achieve our goals, UniCarbKB is partnering with BCSDDB (Bacterial Carbohydrate Structure Database), GlycomeDB, GLYCOSCIENCES.de JCGGDB (Japan Consortium for Glycobiology and Glycotechnology Database), MonosaccharideDB to develop a standard Resource Description Framework RDF representation for carbohydrate structure, biological and bibliographic annotations and experimental evidence. Access to data stored in this format will allow users to perform queries that were not previously possible, and provide the ideal platform for connecting these disparate resources.

While we are still in the early development phases, we have designed a scalable web-friendly framework that integrates information from GlycoSuiteDB and EUROCarbDB. UniCarbKB is a representation of the tremendous growth in information available in glycomics and the adoption of leading-edge technologies to disseminate and query this knowledgebase.

UniCarbKB is based on the reengineering of GlycoSuiteDB and EUROCarbDB and built on the foundations of lightweight Java Rails architecture implementing new search features to explore the wealth of new data now available. The new version will be on-line late 2012. The framework adopts agreed standards to store structural and metadata content including the translation of GlycoSuiteDB structure entries into the GlycoCT format offering a comprehensive structure database (Herget et al., 2008). Significant improvements to the data schema have enabled the merger of these two databases in particular the rational adoption of taxonomic, tissue and disease ontologies. The schema is module in design to segregate the three components (i) structure (ii) informative metadata and (iii) supporting analytical data.

Results and Discussion

New information relevant to glycoproteins, notably the inclusion of glycosylated structures local-

ised in different tissues sourced from a literature exploration study was incorporated. This led to build an accessible database of qualitative and quantitative protein glycoprofilng data. In parallel, special effort is invested into linking this information with sugar recognition curated data (e.g., SugarBind and CFG Glycan Array) to allow deeper mining of the functional role of glycans. At this stage, our first focus is on infectious diseases.

The overall aim of the project is to access, query and mine the most comprehensive bio-curved overview of existing glycoinformation associated with proteins in a site-specific manner both from the attachment and the recognition perspective.

Acknowledgements

The initiative is supported by NECTAR (Australian National eResearch Collaboration Tools and

Resources), STINT (Swedish Foundation for International Cooperation in Research and Higher Education) and SNSF (Swiss National Science Foundation).

References

1. Campbell MP et al. UniCarbKB: putting the pieces together for glycomics research. *Proteomics* (2011) 11(21): 4117-21.
2. UniCarbKB: <http://unicarbkb.org/>
3. GlycosuiteDB: <http://glycosuitedb.expasy.org/glycosuite/glycodb>
4. EUROCarbDB: <http://www.eurocarbdb.org/>
5. SugarBind: <http://sugarbind.expasy.org/sugarbind/>
6. CFG Glycan Array: <http://www.functionalglycomics.org/glycomics/publicdata/primaryscreen.jsp>
7. Herget S, Ranzinger R, Maass K, Lieth CW. GlycoCT-a unifying sequence format for carbohydrates. *Carbohydr Res.* (2008) 343(12):2162-71.