Beilstein-Institut
Cracking the Sugar Code by Navigating the Glycospace
June 27th – July 1st, 2011, Potsdam, Germany

Bioinformatics Analysis of the Glycome Guides Automated Oligosaccharide Synthesis

Daniel Kolarich[1] and Peter H. Seeberger[1,2,3,*]

[1]Max-Planck Institute for Colloids and Interfaces, Am Mühlenberg 1, 14476 Potsdam, Germany,

[2]Free University of Berlin, Arnimallee 22 14195 Berlin, Germany and

[3]The Burnham Institute, La Jolla, CA, U.S.A.

E-Mail: *peter.seeberger@mpikg.mpg.de

Received: 6th January 2012 / Published: 11th July 2012

Deciphering the Glycocode

MIT Technology Review considers glycomics to be one of the ten technologies with the capacity to change the world [1]. Every class of major biomolecule identified so far either has a carbohydrate as a major constituent (e. g. 2-deoxyribose in DNA) or occurs also in a glycosylated form. Glycoproteins and glycolipids are key molecules involved in cell-cell interaction or cell-signalling, proteoglycans and glucosaminoglycans are crucial components of the animal extracellular matrix whereas certain carbohydrate polymers are important energy storage molecules [2, 3]. With the exception of DNA, the biosynthesis of the carbohydrate portion of biomolecules is not a template driven process but the result from concerted actions of numerous glycosyltransferases and glycosyl-donor synthesizing enzymes [2]. These biosynthetic pathways provide the cell with the possibility to fine-tune particular features of glycosylated biomolecules without modifying the actual activity. The capacity to completely reverse IgG function has been shown for minor modifications in the IgG N-glycan structure. The addition of a single sialic acid molecule is able to convert IgG from being a pro-inflammatory into an anti-inflammatory agent [4] and the addition of a core fucose residue the very same N-glycan can inhibit initiation of antibody-dependent cell cytotoxicity by obstructing its binding to the FcγRIIIa receptor [5] without actually changing the actual binding properties towards its antigen.

In contrast to protein sequences, glycan structures cannot be predicted from a template, which makes robust and solid methods for determination of the glycome a central necessity. To date, no methods are available that allow capturing the entire glycome. Due to the vast diversity of different glycosylated molecules, the lack of adequate bioinformatic tools and available satisfactory data repositories, as they are known for genomics and proteomics, it remains questionable whether this goal can be achieved in the near term. Thus, current glycomics approaches naturally focus on a particular class of glycosylated molecules like protein-bound glycans, glycolipids or proteoglycans, to name a few. Due to their diverse nature, different approaches are required in order to enable their analysis, characterisation and sequencing. Nevertheless, many of the glycosylated biomolecules are just present in relative low amounts, and targeted purification for studying their biological role is often impossible. A comprehensive automated synthesis allows access to larger quantities of defined glycans of biological relevance [6, 7]. This capacity provides central means required in deciphering the glyco-code of glycan sequences.

Decomplexing the Glycome for Automated Synthesis

The biosynthetic nature of the glycome provides the cell with the capacity to store and transmit exponentially more information in glycan sequences compared to the genome or proteome (Table 1). Compared to the linear structure of DNA or protein sequences, glycans can occur in branched structures. Furthermore the different monosaccharide building blocks can be linked together in different ways, resulting in substantially different biomolecules of similar sequence but different biological properties. This is best illustrated by comparing starch and cellulose: both polymers consist of 1 – 4 linked glucose monosaccharide building blocks, but differ in the stereochemistry of the linkage: α-1,4 linkages dominate in starch whereas β-1,4 linkages are the major type of connection in cellulose, resulting in two very different types of biopolymers. Similarly, comparably small differences in glycosylation and linkage can have significant impact on the biological activity of mammalian biomolecules as pointed out above.

Despite the immense variety of theoretically possibly conformations a hexasaccharide could have (Table 1), a substantially smaller number of different types of linkages are found in mammalian oligosaccharides [8]. Using the database information stored in GLYCOSCEINCES.de database (www.glycosciences.de/) the structural diversity of mammalian carbohydrates was explored and the most common monosaccharide building blocks found in mammalian carbohydrates were identified [8]. About three quarters of structures found in the database show some type of branching at least once. Three monosaccharide building blocks, namely glucosamine, galactose, and mannose contribute to about 75% of all building blocks found, and the major terminating residues comprise alpha-linked sialic acid, alpha-linked fucose, and beta-linked galactose. Interestingly, glucose, which is from the quantitative point

of view the most abundant monosaccharide found in mammals, just plays a very minor role in glycoconjugate glycan structures. This data indicates that there is a clear division in storage and glycoconjugate monosaccharides.

Table 1. Number of theoretically possibly structures of biopolymers (Reprinted with permission from Werz *et al* [8]. Copyright 2007 American Chemical Society").

Oligomer size	Numbers of different oligomers		
	Nucleotides	Peptides	Carbohydrates
1	4	20	20
2	16	400	1 360
3	64	8 000	126 080
4	256	160 000	13 495 040
5	1 024	3 200 000	1 569 745 920
6	4 096	64 000 000	192 780 943 360

In the context of automated synthesis the most important result is, however, the finding that a manageable number of just 36 building blocks can be used to construct about 75% of the structures available in the database (Figure 1) [8]. This encouraging data demonstrates that automated synthesis has the capacity to produce significant amounts highly defined glycan structures that enable glycobiologists to decipher the messages hidden within mammalian glycan structures.

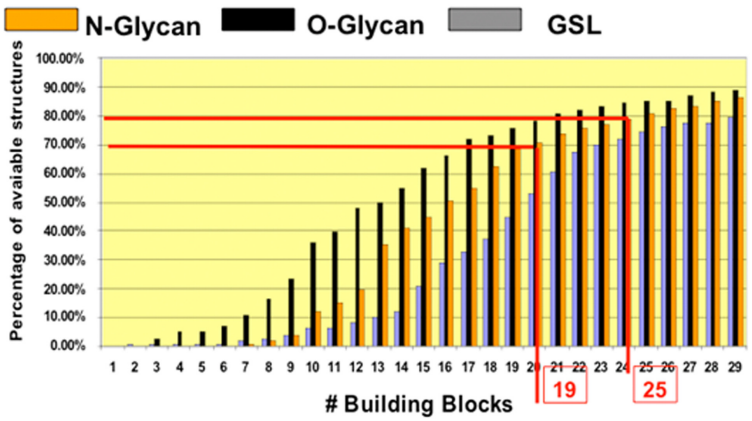

Figure 1. Number of building blocks required for synthetic access to mammalian carbohydrates. Percentage of accessible mammalian carbohydrates correlated to the number of building blocks in the context of the different classes analysed (glycolipids and N- and O-linked glycans) and correlated to the number of building blocks. (Reprinted with permission from Werz *et al.* [8]. Copyright 2007 American Chemical Society").

Similar encouraging results were obtained using a bioinformatics assessment of monosaccharide building blocks present in bacterial sugars [9]. Data on bacterial sugars deposited in Bacterial Carbohydrate Structure Database (BCSDB) (csdb.glycoscience.ru/bacterial/) and GLYCOSCIENCES.de (GS) has been statistically evaluated focusing on pathogenic bacteria with regard to five major parameters:

- monosaccharide units abundance,
- disaccharide pairs,
- carbohydrate modifications,
- presence and use of sialic acids and
- class-specific monosaccharides.

Not surprisingly, bacteria showed to be more diverse in the monosaccharide building blocks, and significant differences in the use of these has been found for different classes of bacterial species [9]. Nevertheless an infinite number of about 25 monosaccharides allows to build up about 71% of the bacterial glycome known so far.

Glyco-bioinformatics – Quo vadis?

The findings of these two straightforward studies point out the fundamental necessity for adequate and openly accessible data repositories. Glyco-bioinformatics has been characterized by the existence of multiple disconnected and incompatible islands of experimental data, data resources and specific applications, managed by various consortia, institutions or local groups. These resources rarely provided the necessary communication mechanisms that would allow for the efficient combination and comparison of these data. However, combinations and comparison of data is a crucial point as it enables researches to perform broader bioinformatic studies in a Systems Biology context on glycoconjugates.

Databases clearly need data, but what type of data needs to be stored? Should quantity be more important than quality of the data? Clearly a highly accurate database consisting of a very limited number of entries will hardly be beneficial, but likewise will a large database with inadequately curated data be of little value. Another issue to be considered is specific for glycoconjugates. In contrast to DNA, which can be easily categorised into different species, this is more challenging for glycoconjugates since similar glycans can occur across different species. User and data friendly solutions have to be developed to address these and many additional challenges that are peculiar for glycoconjugates.

A promising new development in this context is UNICARB-DB (www.unicarb-db.org) [10] and for more detailed information the contribution from Packer *et al*. in this volume of the proceedings should be consulted [11]. Equally important in this context is also the information on how the data in glycomics experiments is obtained. Particular reporting guidelines

are in place in the related fields of genomics and proteomics, [12, 13], and recently a similar initiative called MIRAGE (**M**inimum **I**nformation **R**equired for **A** **G**lycomics **E**xperiment, initiated by Prof. Will York) has formed out of this 2nd Beilstein Symposium on Glyco-bioinformatics. For more information on that the contribution from Prof. York in this proceedings issue is referred to [14].

Conclusions

Glycomics research urgently requires robust, reliable and accessible glyco-bioinformatics solutions that can be utilised by researches from the many diverse fields closely related to the synthesis, analysis and biological investigation of glycoconjugates. The developments in robust and user-friendly databases, the constant deposition of newly acquired data and novel bioinformatic tools will boost subsequent research in glycoconjugate systems biology. These developments will provide a solid basis that will allow researchers to understand the diverse interaction networks glycoconjugates are involved in. For that to happen, the joint efforts initiated in recent years and also within the course of the 2nd Beilstein Symposium on Glyco-bioinformatics need to be continued.

References

[1] Negroponte, N. (2003) Creating a Culture of Ideas. *Technology Review*. **106**(1): 34 – 49.

[2] Varki, A., *et al.* (2009) *In:* Essentials of Glycobiology, 2nd edition. New York: Cold Spring Harbor Laboratory Press. 784.

[3] Seeberger, P.H. (2005) Exploring life's sweet spot. *Nature* **437**(7063):1239. doi: http://dx.doi.org/10.1038/4371239a.

[4] Kaneko, Y., Nimmerjahn, F. and Ravetch, J.V. (2006) Anti-inflammatory activity of immunoglobulin G resulting from Fc sialylation. *Science*. **313**(5787):670 – 3. doi: http://dx.doi.org/10.1126/science.1129594.

[5] Iida, S., *et al.* (2006) Nonfucosylated therapeutic IgG1 antibody can evade the inhibitory effect of serum immunoglobulin G on antibody-dependent cellular cytotoxicity through its high binding to FcgammaRIIIa. *Clinical cancer research*: an official journal of the American Association for Cancer Research **12**(9): 2879 – 87.

[6] Seeberger, P.H. and Werz, D.B. (2007) Synthesis and medical applications of oligosaccharides. *Nature* **446**(7139):1046 – 51. doi: http://dx.doi.org/10.1038/nature05819.

[7] Seeberger, P.H. (2009) Chemical glycobiology: why now? *Nat. Chem. Biol.* **5**(6):368 – 72.
doi: http://dx.doi.org/10.1038/nchembio0609-368.

[8] Werz, D.B., *et al.* (2007) Exploring the structural diversity of mammalian carbohydrates (“glycospace”) by statistical databank analysis. *ACS chemical biology* **2**(10):685 – 91.
doi: http://dx.doi.org/10.1021/cb700178s.

[9] Adibekian, A., *et al.* (2011) Comparative bioinformatics analysis of the mammalian and bacterial glycomes. *Chemical Science* **2**(2):337 – 344.
doi: http://dx.doi.org/10.1039/c0sc00322k.

[10] Campbell, M.P., *et al.* (2011) UniCarbKB: Putting the pieces together for glycomics research. *Proteomics* **11**(21):4117 – 21.
doi: http://dx.doi.org/10.1002/pmic.201100302.

[11] Campbell, M.P., Lisacek, F., Wilkins, M.R., Rudd, P.M., Kolarich, D., Hayes, C.A., Karlsson, N.G. and Packer, N.H. (2012) Linking Glycomics Repositories with Data Capture. *In*: Proceedings of 2nd the Beilstein Symposium on Glyco-Bioinformatics (Eds. M.G. Hicks and C. Kettner), Logos-Verlag, Berlin, pp 165–178.

[12] Field, D., *et al.* (2008) The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol.* **26**(5):541 – 7.
doi: http://dx.doi.org/10.1038/nbt1360.

[13] Taylor, C.F., *et al.* (2007) The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* **25**(8):887 – 93.
doi: http://dx.doi.org/10.1038/nbt1329.

[14] York, W. and Ranzinger, R. (2012) MIRAGE: Minimum Information Required for a Glycomics Experiment. *In*: Proceedings of the 2nd Beilstein Symposium on Glyco-Bioinformatics (Eds. M.G. Hicks and C. Kettner), Logos-Verlag, Berlin, pp 29–38.