Systems analysis of bacterial glycomes

Emily Kay*†, Victor I. Lesk*, Alireza Tamaddoni-Nezhad*, Paul G. Hitchen*, Anne Dell*, Michael J. Sternberg*, Stephen Muggleton* and Brendan W. Wren*†[1]

*The Centre for Integrative Systems Biology, Imperial College, London SW7 2AZ, U.K., and †Department of Infectious and Tropical Diseases, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, U.K.

Abstract

Bacteria produce an array of glycan-based structures including capsules, lipo-oligosaccharide and glycosylated proteins, which are invariably cell-surface-located. For pathogenic bacteria, such structures are involved in diverse roles in the life cycle of the bacterium, including adhesion, colonization, avoidance of predation and interactions with the immune system. Compared with eukaryotes, bacteria produce huge combinatorial variations of glycan structures, which, coupled to the lack of genetic data, has previously hampered studies on bacterial glycans and their role in survival and pathogenesis. The advent of genomics in tandem with rapid technological improvements in MS analysis has opened a new era in bacterial glycomics. This has resulted in a rich source of novel glycan structures and new possibilities for glycoprospecting and glycoengineering. However, assigning genetic information in predicted glycan biosynthetic pathways to the overall structural information is complex. Bioinformatic analysis is required, linked to systematic mutagenesis and functional analysis of individual genes, often from diverse biosynthetic pathways. This must then be related back to structural analysis from MS or NMR spectroscopy. To aid in this process, systems level analysis of the multiple datasets can be used to make predictions of gene function that can then be confirmed experimentally. The present paper exemplifies these advances with reference to the major gastrointestinal pathogen *Campylobacter jejuni*.

Bacterial glycostructures: important, complex and poorly understood

The bacterial glycome represents the total set of glycans expressed by a bacterial cell. Glycans serve diverse functions in bacterial cells and are often found attached to proteins (glycoproteins) and lipids (glycolipids) or are loosely associated as cell-surface polysaccharides. The varied glycostructures are found in different types of bacteria [1]. For example, in Gram-negative bacteria, these include LPS (lipopolysaccharide), LOS (lipo-oligosaccharide), peptidoglycan and CPS (capsular polysaccharide); in Gram-positive bacteria, these include extrapolysaccharide and lipoteichoic acid; and in specialized bacteria, such as mycobacteria, these include mycolic acid, arabinogalactan and lipoarabinomannan. In addition, both O- and N-linked protein glycosylation systems are increasingly found in bacteria that contribute to the rich sugar decoration of surface structures such as pili and flagella. Given that the glycan decoration of a bacterial cell surface is often the first information about the cell presented to the outside world, it is unsurprising that such diverse surface structures are important in recognition processes between the bacterium and the host or environment. This translates to vital roles in the life cycle of bacteria, including symbiosis, pathogenesis, cell–cell interactions, predation avoidance and immune evasion. Deconvoluting the importance of each glycan in these processes is challenging owing to many factors, not least the complexity of the glycans.

Bacterial glycan structures are highly complex and are manifest in vast combinations. In contrast with mammalian glycans, which rely on a core group of a dozen common monosaccharide building block units, the potential number of bacterial equivalents is indeterminate, particularly as, currently, only approx. 0.1% of all known bacteria can be cultured and characterized. In addition to the mammalian monosaccharide units, pentoses, heptoses and nonuloses are also often found in bacteria. To add further variation, the monosaccharide building blocks assemble as linear and/or branched arrangements coupled to further complexity through regiochemistry and stereochemistry of the anomeric carbon in the glycosidic bond. The complexity can be compounded further by the different modes of attachment, including glycolipids and serine/threonine (O-) and asparagine (N-) linkages to proteins. Surface glycostructures may be actively varied by the bacteria, e.g. phase-variation in *Campylobacter* LOS leading to alternative presentation of ganglioside mimics [2].

The sheer variety of unusual sugars and linkage diversity presents the cell with much opportunity to carry recognition information in cell–cell interactions; however, it also makes bacterial glycans difficult to study. Although there have been considerable advances in analytical techniques such as MS and NMR spectroscopy coupled with developments in genomics, bioinformatics, metabolomics and lipidomics, data are being

Key words: bioinformatics, *Campylobacter jejuni*, glycan biosynthesis, glycome, Inductive Logic Programming (ILP), systems microbiology.

Abbreviations used: CPS, capsular polysaccharide; ILP, Inductive Logic Programming; LOS, lipo-oligosaccharide.

[1]To whom correspondence should be addressed (email Brendan.Wren@lshtm.ac.uk).

generated much faster than they can be effectively analysed. Lectin microarrays have recently come to the fore in sampling carbohydrate-based bacterial diversity, allowing real-time monitoring of changes in glycan repertoire in response to changing conditions [3]. However, these techniques are limited to commercially available lectins. The challenge now is how to weave the different strands of information to provide a holistic understanding of the roles that glycans play in the bacterial cell. In contrast with genomics and proteomics that are accessible to most scientists, glycomics is still at a developmental stage. A systems approach to unravelling bacterial glycomes offers much hope, but is still in its infancy.

Systems microbiology

Systems biology aims to study the dynamic interactions of more than one component in a system in order to understand and predict the behaviour of the system as a whole. Typical approaches involve an iterative cycle of 'dry lab' modelling and 'wet lab' verification. A system can be a single cell, an organ, an organism or an environment containing interacting organisms. For microbes, this would usually apply to a single cell that has led to the new field of systems microbiology [4]. The genomic era, particularly for microbes, has given rise to advancements in genome-level studies of DNA content, transcriptomics, proteomics and metabolomics. In the context of bacterial glycomics, systems biology is used to understand fully the biosynthetic pathways required for glycan synthesis, structure and display. We are now in the position of having large amounts of data about biological systems, but the challenge is in being able to integrate and model multiplatform datasets in order to gain novel biological insight. In the present article, we review one such approach using logic-based machine learning to interpret glycan biosynthesis in the major gastrointestinal pathogen *Campylobacter jejuni*.

The *C. jejuni* NCTC11168 glycome

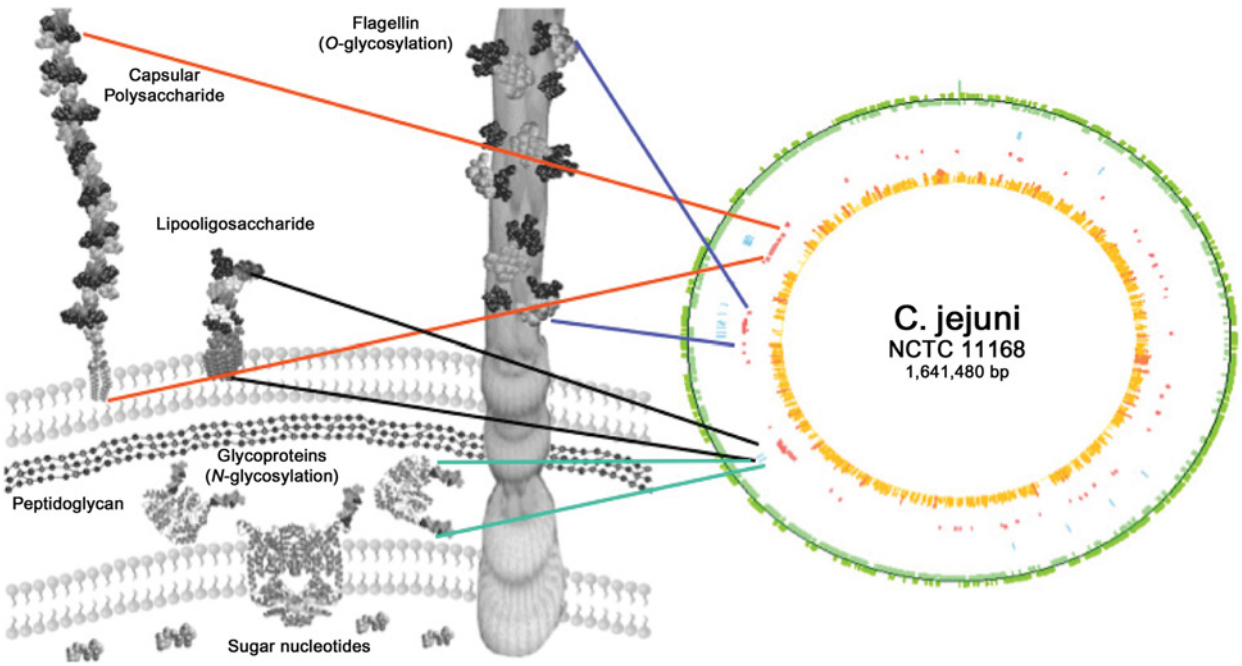
C. jejuni is a Gram-negative flagellated spiral bacterium, and is a leading cause of human gastrointestinal disease worldwide. A rare, but serious, post-immune complication is Guillain–Barré syndrome, which is associated with molecular mimicry, whereby the bacterial LOS mimics the carbohydrate structure of gangliosides on human nerve tissue, leading to paralysis via the production of autoimmune antibodies. *C. jejuni* has at least four glycostructures important in pathogenesis, avoiding bacteriophage predation and immune evasion [5,6]. These include the LOS, the CPS and both N- and O- linked glycosylation pathways, the latter of which decorates the flagellum (Figure 1). Since the first description of the genome sequence of the *C. jejuni* strain NCTC 11168 in 2000 [7], re-annotated in 2007 [8], several attempts have been made to relate gene sequence to glycan structure and function [6]. Unravelling the biosynthetic pathways of these glycostructures is complicated, not least because there is a degree of cross-talk between the pathways and also enzymes from the different pathways may be able to compensate for enzymes mutated in another pathway, e.g. Gne, a bifunctional UDP-GlcNAc/Glc 4-epimerase involved in the LOS and the N- and O- linked glycosyation pathways [9], and PseB from the O-linked pathway is able to perform a secondary reaction to give the same product as PglF from the N-linked pathway [10]. Understanding how the glycan structures are synthesized and how these may alter under different biologically relevant conditions may shed light on the environmental survival and pathogenesis of this fastidious human pathogen. Glycans have already been shown to play an important role in host colonization [11], and it has been suggested that the phosphoramidate capsule modification is more likely to be associated with severe disease outcome [12]. This question is ideally suited to a systems biology modelling approach. Relating the presence of genes to the overall structure is a continuing goal. Some gene products have been definitively assigned to specific enzymes, but most, particularly for the CPS, remain obscure in part due to novel functionality.

Logic-based machine learning

The approach we have used to unravel biosynthetic pathways is a logic-based machine learning approach termed ILP (Inductive Logic Programming) [13]. ILP is a combination of logic programming and machine learning and aims to use background knowledge to generate a hypothesis that is consistent both with the background knowledge and observables (experimental data). In this instance, background knowledge is derived from metabolic networks, glycan structures and enzymatic functions from publically available databases, and converted into the programming language prolog. This can be achieved using the data integration platform ONDEX [14]. Capsule phenotype determined by HR-MAS (high-resolution magic-angle spinning) NMR data from a set of mutants in the CPS locus of *C. jejuni* NCTC 11168 was also used to inform the model. From the background knowledge, program rules are generated: ILP has been used to teach computers to play chess, as only certain moves are possible by certain pieces and, in the same way, only certain classes of enzymes are capable of carrying out specific reactions within a biochemical pathway. Using this approach, it is possible that more than one hypothesis may be generated that fits the background knowledge and experimental observations equally well. Some constraint is needed in order to select the most likely hypothesis. This is done by finding the most specific hypothesis, i.e. the hypothesis that explains the greatest number of variables. If you imagine a jigsaw puzzle with several pieces missing, but pretend that the jigsaw is complete except for one piece, then try all the lone pieces in that gap, repeat this for all gaps, then come up with a probability weighting of which piece is most likely to fit which gap. In order to test the predictive accuracy of the model, a training example is used. In this case, the *C. jejuni* NCTC 11168 LOS pathway was chosen as most of the 21 genes in the locus have been related to enzymes that

Figure 1 | From genome sequence to glycostructure

The outermost concentric circle of bars represents the 1654 predicted coding sequencing in *C. jejuni* NCTC 11168, the next set of bars are predicted phase-variable genes, and the next inner ring of bars are predicted surface structures, of which some encode the CPS, LOS and O- and N-linked gycosylation pathways (the inner circle of bars are *Helicobacter pylori* orthologues). Adapted with kind permission from Spring Science+Business Media: In Bacterial Genomes and Infectious Diseases (V.L. Chan, P.M. Sherman and B. Bourke, eds), *Campylobacter*: From Glycome to Pathogenesis (2006), pp. 63–90, J. Kelly, J.-R. Brisson, N.M. Young, H.C. Jarrell and C.M. Szymanski, Figure 2, and [7] with permission.



contribute to the overall glycostructure. Successive amounts of background knowledge are then withheld from the model in order to see whether the pathway can be accurately reconstituted both before and after rounds of learning. The background knowledge withheld can include enzymes that code for certain reactions as well as reactions themselves, as the model should be able to predict unknown reactions as well as the enzymes responsible for catalysing them. Evidently the more background knowledge you have, the fewer unknowns and the greater predictive accuracy of the model, which is why successive rounds of experiments are often needed to refine the model, especially with bacterial glycostructures, as background knowledge on pathways and enzymatic reactions is sparse. The ILP approach has been used successfully for a 'robot scientist' that was able to select and carry out experiments to reconstruct the amino acid synthesis pathway of *Saccharomyces cerevisiae* [15].

Applying the model to the CPS system with relevant mutagenesis/structural data, the learned model suggested explicit hypotheses concerning the functions of uncharacterized gene products. Some of these hypotheses are intuitive, as mutation leads to loss of a specific capsule moiety; others are less intuitive, as mutation leads to an acapsular phenotype. These include predictions about the involvement of some putative transferases of unknown function in the assembly of CPS. Currently, these predictions of novel functions are being experimentally verified.

Conclusions

We have probably only scratched the surface in terms of understanding the importance and utility of bacterial glycans. They are not only vital in the survival and virulence of bacteria, but also important targets for antimicrobial and vaccine design. Furthermore, the demonstration that bacterial oligosaccharyltransferases (first discovered in *C. jejuni*) can be used to produce recombinant glycoproteins in *Escherichia coli*, promises a new era in glycoengineering [16,17]. This will be fuelled by the diversity of bacterial glycan structures and oligosaccharyltransferases that are continually being discovered and will provide the glycotoolbox for recombinant glycobiology. Synthetic biology that depends on a registry of standard biological parts (http://parts.mit.edu) and new vectors for the assembly and propagation of cells will benefit from a well-defined glyco parts list. Additionally, the protein glycosylation pathways in simple bacterial cells are useful models for the more complex eukaryotic counterparts [18]. However, before this potential can be realized, a systematic and predictable understanding of glycan

biosynthetic pathways, from gene content to structure and function, is required.

We have initiated attempts to use inductive logic programming to model the biosynthesis of *C. jejuni* glycans, with promising initial results. A key advantage of the logical modelling approach over other approaches such as Bayesian networks is to be able to incorporate background knowledge of existing known biochemical pathways, together with information on enzyme classes and reaction chemistry. With the increasing emphasis on high-throughput experimental initiatives, machine learning and other automated methods of analysis will become increasingly important for many biological problems. Systems glycobiology is clearly a multidisciplinary research area on the move. With the application of further mathematical, bioinformatic and modelling approaches on bacterial glycans, an Aladdin's cave of riches is waiting to be mined.

Acknowledgements

We acknowledge colleagues at the Centre for Integrated Systems Biology at Imperial College for invaluable discussions

Funding

This work is funded by the Biotechnology and Biological Sciences Research Council [grant number BB/C519670/1].

References

1 Reid, C.W., Fulton, K.M. and Twine, S.M. (2010) Never take candy from a stranger: the role of the bacterial glycome in host-pathogen interactions. Fut. Microbiol. **5**, 267–288
2 Guerry, P., Szymanski, C.M., Prendergast, M.M., Hickey, T.E., Ewing, C.P., Pattarini, D.L. and Moran, A.P. (2002) Phase variation of *Campylobacter jejuni* 81–176 lipooligosaccharide affects ganglioside mimicry and invasiveness *in vitro*. Infect. Immun. **70**, 787–793
3 Hsu, K.-L., Pilobello, K.T. and Mahal, L.K. (2006) Analyzing the dynamic bacterial glycome with a lectin microarray approach. Nat. Chem. Biol. **2**, 153–157
4 Kay, E. and Wren, B.W. (2009) Recent advances in systems microbiology. Curr. Opin. Microbiol. **12**, 577–581
5 Coward, C., Grant, A.J., Swift, C., Philp, J., Towler, R., Heydarian, M., Frost, J.A. and Maskell, D.J. (2006) Phase-variable surface structures are required for infection of *Campylobacter jejuni* by bacteriophages. Appl. Environ. Microbiol. **72**, 4638–4647
6 Karlyshev, A.V., Ketley, J.M. and Wren, B.W. (2005) The *Campylobacter jejuni* glycome. FEMS Microbiol. Rev. **29**, 377–390
7 Parkhill, J., Wren, B.W., Mungall, K., Ketley, J.M., Churcher, C., Basham, D., Chillingworth, T., Davies, R.M., Feltwell, T., Holroyd, S. et al. (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. Nature **403**, 665–668
8 Gundogdu, O., Bentley, S.D., Holden, M.T., Parkhill, J., Dorrell, N. and Wren, B.W. (2007) Re-annotation and re-analysis of the *Campylobacter jejuni* NCTC11168 genome sequence. BMC Genomics **8**, 162
9 Bernatchez, S., Szymanski, C.M., Ishiyama, N., Li, J., Jarrell, H.C., Lau, P.C., Berghuis, A.M., Young, N.M. and Wakarchuk, W.W. (2005) A single bifunctional UDP-GlcNAc/Glc 4-epimerase supports the synthesis of three cell surface glycoconjugates in *Campylobacter jejuni*. J. Biol. Chem. **280**, 4792–4802
10 Guerry, P., Ewing, C.P., Schoenhofen, I.C. and Logan, S.M. (2007) Protein glycosylation in *Campylobacter jejuni*: partial suppression of pglF by mutation of pseC. J. Bacteriol. **189**, 6731–6733
11 Howard, S.L., Jagannathan, A., Soo, E.C., Hui, J.P.M., Aubry, A.J., Ahmed, I., Karlyshev, A., Kelly, J.F., Jones, M.A., Stevens, M.P. et al. (2009) *Campylobacter jejuni* glycosylation island important in cell charge, legionaminic acid biosynthesis, and colonization of chickens. Infect. Immun. **77**, 2544–2556
12 McNally, D.J., Lamoureux, M.P., Karlyshev, A.V., Fiori, L.M., Li, J., Thacker, G., Coleman, R.A., Khieu, N.H., Wren, B.W., Brisson, J.-R. et al. (2007) Commonality and biosynthesis of the O-methyl phosphoramidate capsule modification in *Campylobacter jejuni*. J. Biol. Chem. **282**, 28566–28576
13 Muggleton, S. (1991) Inductive logic programming. New Gener. Comput. **8**, 295–318
14 Kohler, J., Baumbach, J., Taubert, J., Specht, M., Skusa, A., Ruegg, A., Rawlings, C., Verrier, P. and Philippi, S. (2006) Graph-based analysis and visualization of experimental results with ONDEX. Bioinformatics **22**, 1383–1390
15 King, R.D., Whelan, K.E., Jones, F.M., Reiser, P.G.K., Bryant, C.H., Muggleton, S.H., Kell, D.B. and Oliver, S.G. (2004) Functional genomic hypothesis generation and experimentation by a robot scientist. Nature **427**, 247–252
16 Wacker, M., Linton, D., Hitchen, P.G., Nita-Lazar, M., Haslam, S.M., North, S.J., Panico, M., Morris, H.R., Dell, A., Wren, B.W. and Aebi, M. (2002) *N*-linked glycosylation in *Campylobacter jejuni* and its functional transfer into *E. coli*. Science **298**, 1790–1793
17 Langdon, R.H., Cuccui, J. and Wren, B.W. (2009) N-linked glycosylation in bacteria: an unexpected application. Fut. Microbiol. **4**, 401–412
18 Szymanski, C.M. and Wren, B.W. (2005) Protein glycosylation in bacterial mucosal pathogens. Nat. Rev. Microbiol. **3**, 225 – 237

Received 12 March 2010
doi:10.1042/BST0381290