Suggested reviewers :

Mario Feldman (Washington University in St Louis, [mariofeldman@wustl.edu](mailto:mariofeldman@wustl.edu))

Kaspar Locher (ETH Zürich, [locher@mol.biol.ethz.ch](mailto:locher@mol.biol.ethz.ch))

Antonio Molinaro (University of Napoli Federico II, [molinaro@unina.it](mailto:molinaro@unina.it))

Chris Whitfield (University of Guelph; [cwhitfie@uoguelph.ca](mailto:cwhitfie@uoguelph.ca))

Jochen Zimmer (University of Virginia, [jz3x@virginia.edu](mailto:jz3x@virginia.edu))

Diversity of sugar-diphospholipid-utilizing glycosyltransferase families

Ida Meitil1, Garry P. Gippert1, Kristian Barrett1, Cameron J. Hunt1, Bernard Henrissat1,2,3∗

1 Department of Biotechnology and Biomedicine, Technical University of Denmark, Kgs. Lyngby, Denmark,

2 Architecture et Fonction des Macromolécules Biologiques, CNRS, Aix-Marseille Université, Marseille, France,

3 Department of Biological Sciences, King Abdulaziz University, Jeddah, Saudi Arabia

**Keywords:** glycosyltransferases; sequence similarity;  glycan similarity; classification; bacterial surface polysaccharides

\*Author for correspondence:

Bernard Henrissat: ORCID number 0000-0002-3434-8588

e-mail: [behen@dtu.dk](mailto:behen@dtu.dk)

**ORCID numbers:**

Ida Meitil: 0000-0001-9214-0980

Garry P. Gippert: 0000-0002-6803-9951

Kristian Barrett: 0000-0002-4170-4253

Cameron J. Hunt: 0000-0002-8495-0542

**Abstract**

Glycosyltransferases, the enzymes that assemble glycans from activated sugar donors, are classified in 116 sequence-based families in the carbohydrate-active enzymes database (CAZy; www.cazy.org). Because these families correlate with enzyme structure, mechanism and some aspects of reaction products, they provide predictive power for inferring enzyme properties from sequence only. Here we expand the CAZy family classification of GT-C glycosyltransferases that utilize an oligosaccharide activated by a diphospholipid. While straightforward for some families, the sequence diversity of bacterial polysaccharide polymerases prevented the creation of a single family. To address this, we employed sequence similarity networks to define groups of alignable sequences, HMMs were built for each group, and the resulting HMMs were aligned to form families, many of which could not be identified by global multiple sequence alignment. In total, we identified 17 families and analyzed the donor and acceptor substrates to assess their correlation with the products' chemical structure. The families exhibited conservation of reaction stereochemistry and structural resemblance of synthesized glycans, a feature for which we developed an original similarity score in the absence of an existing scoring system. Additionally, distant interfamily relationships that shared the stereochemistry of the created glycosidic bond, were used to define ‘clans’ of related families.