**New abstract**

Bacterial glycosyltransferases (GTs) are essential for decorating the cell surface with polysaccharides that define the phenotype, and specifically those four defined as peptidoglycan polymerases, enterobacterial common antigen polymerases, O-antigen ligases, and other bacterial polysaccharide polymerases (BP-Pol) are not well understood, but highly decisive for pathogenicity. In this study we have defined 17 new GT families in the carbohydrate-active enzymes database (CAZy; [www.cazy.org](http://www.cazy.org)), which all share the particularity of using diphospholipid-activated sugars and which were previously missing in CAZy.

a. CAZy did not previously cover a distinct subclass of bacterial glycosyltransferases, called GT-CB, which are all membrane proteins with 10 transmembrane helices.  
b. This paper attempts to expand the number of GT-CB families in CAZy.  
c. GT-CBs fall into four large subgroups of functions, peptidoglycan polymerases, enterobacterial common antigen polymerases (ECA-Pol), O-antigen ligases (O-Lig), and other bacterial polysaccharide polymerases (BP-Pol).  
d. The first three subgroups form well defined families of proteins that are similar enough in sequence be aligned in a multiple sequence alignment, while the BP-Pol group is highly diverse in sequence, and cannot be aligned.  
e. A strategy was developed to identify families within the BP-Pol group, which involved manual review and adjustment of sequence similarity networks (SSNs) to define small families of highly similar sequences, followed by using HMM comparisons to merge small families together into a single, larger family when they have very similar amino acid sequence profiles.  
f. Using this strategy, the authors created 14 families of BP-Pols, which cover approximately 40% of the BP-Pols of known substrate specificity. The remainder form small groups which may be expanded and added to CAZy in the future as more bacterial genomes are sequenced.  
g. An important advance is that the authors identify four distinct subgroups within GT-CB, each of which conserves the overall reaction product stereochemistry.

Bacterial surface polysaccharides are essential for bacterial function and defining for pathogenicity. Yet, the glycosyltransferases (GT) that build them are not well understood and were previously missing in the carbohydrate-active enzymes database (CAZy; [www.cazy.org](http://www.cazy.org)). This particularly relates to four functional classes; peptidoglycan polymerases, enterobacterial common antigen polymerases, O-antigen ligases, and other bacterial polysaccharide polymerases (BP-Pol), which all share the particularity of using diphospholipid-activated sugars. In this study we have defined 17 new GT families covering these four functionalities.

While the first three functional classes formed well-defined families of similar proteins, the sequences of the bacterial polysaccharide polymerases were highly diverse, which prevented the creation of a single family. To address this, we developed a new clustering method where a sequence similarity network was used to define groups of alignable sequences, hidden Markov models (HMMs) were built for each group, and the resulting HMMs were aligned to form new families. This approach appeared useful for clustering of highly diverse sequences compared to phylogenetic analyses which rely only on global multiple sequence alignments.

Furthermore, we combined the protein sequence information with known data about surface polysaccharide structures and found that each of the new BP-Pol families, appear to retain the overall reaction stereochemistry of the product and are specific towards the transferred sugars.

On a higher level, comparison of the new GTs revealed three clans of distantly related families with the same stereochemical mechanism, and we identified potential critical amino acids for catalysis in the BP-Pols.

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.