Cover letter

Kgs. Lyngby, August xx, 2023

Dear Editors of Nature Communications

We are submitting a manuscript entitled “Analyzing the diversity of sugar-diphospholipid-utilizing glycosyltransferase families” for evaluation by *Nature Communications*.

Glycosyltransferases (and other types of carbohydrates-processing enzymes) are classified in sequence-based families in the carbohydrate-active enzymes database (CAZy; www.cazy.org), a widely recognized resource, central to the field, created and updated since 1998 (our most recent paper on this database: Nucleic Acids Res. 2022 Jan 7;50(D1):D571-D577. doi: 10.1093/nar/gkab1045). These families have proved useful as they correlate with enzyme structure and mechanism, while grouping together enzymes that act on different substrates and thus (i) illustrate how they have evolved to acquire different specificities and (ii) provide clues and limits to precise function assignment by sequence similarity. The CAZy database currently lists 116 families of glycosyltransferases (GTs), mostly utilizing carbohydrates activated by nucleotide diphosphates, nucleotide monophosphates and lipid monophosphates. Only one CAZy family currently contains GTs that utilize an oligosaccharide activated by a diphospho-lipid. Thus a large number of GTs involved in bacterial polysaccharide assembly, also utilizing oligosaccharide donors activated by diphospho-lipids, are currently absent from the classification.

In the work reported in our manuscript, we attempted to assemble these missing GTs into CAZy families. Whilst this was rather straightforward for several types of enzymes (peptidoglycan polymerases, O-antigen ligases and Enterobacterial Common Antigen polymerases), the exceptional sequence diversity of other bacterial polysaccharide polymerases (BP-Pol) made it impossible to build a meaningful global multiple sequence alignment that would correspond to one family, a limitation that has so far prevented a deep evaluation of the relationships between the sequence (hence the structure) of the enzymes and their substrate specificity. We have tackled this problem by grouping clearly related and alignable sequences, building HMMs for each group in order to capture the amino acid profile of each group, and then comparing the resulting HMMs to identify relatedness that could not be identified by global multiple sequence alignments. The largest of the groups we have identified define 17 novel GT families in the CAZy database (the families are provisionally termed GTxx1 to GTx17 in the submitted manuscript and will receive final CAZy family numbers at proof stage). Having defined these families, we proceeded to examine the substrates (carbohydrate donor and acceptor) for the bacterial polysaccharide polymerases in order to determine whether the families correlated to the chemical structure of the products. We found that the families display conservation of the stereochemistry of the reaction at the catalytic site and a structural resemblance of the synthesized glycans, for which we had to develop an original similarity score in absence of such a scoring system in the literature. The GT families described here also reveal that the invariant residues in the families not only co-localize in the same area of the three-dimensional structures (whether actual or predicted), but also correspond to the residues found essential for function in the families where this has been studied experimentally. Finally, we report ultra-distant relatedness across several of the newly defined families, insufficient for merging them into single families, but bringing evidence for common ancestry and mechanism. We discuss these distant interfamily relationships, which are also accompanied by conservation of the stereochemistry of the created glycosidic bond, define

'clans' of related families and place these clans in the framework of structural groupings proposed in the literature.

We believe that the addition of 17 novel families of glycosyltransferases (with many more to come as sequence number and diversity increase), the establishment of mechanistic conservation within the families, the emergence of clans of related families, and the description of an algorithm that quantifies glycan similarity, constitute a very significant addition to the literature on this large group of proteins characterized by a sequence diversity such that has hitherto proven impossible to analyze globally. A list of potential reviewers is suggested below.

Sincerely yours

Bernard Henrissat
Professor
The Technical University of Denmark
DTU Bioengineering
Søltofts Plads
2800 Kgs. Lyngby, Denmark
bernard.henrissat@gmail.com or behen@dtu.dk

Suggested reviewers :

Mario Feldman (Washington University in St Louis, mariofeldman@wustl.edu)
Kaspar Locher (ETH Zürich, locher@mol.biol.ethz.ch)
Antonio Molinaro (University of Napoli Federico II, molinaro@unina.it)
Chris Whitfield (University of Guelph; cwhitfie@uoguelph.ca)
Jochen Zimmer (University of Virginia, jz3x@virginia.edu)