Annotation and prediction of carbohydrate binding sites on protein surface

GHEERAERT Aria[A], BAILLY Thomas[A], REN Yani[A,B], HAMRAOUI Ali[A,C], TE Julie[A], VANDER MEERSCHE Yann[A], CRETIN Gabriel[A], LEON FOUN LIN Ravy[A], GELLY Jean-Christophe[A], PEREZ Serge[D], GUYON Frédéric[A] and GALOCHKINA Tatiana[A]

A) Université Paris Cité and Université des Antilles and Université de la Réunion, INSERM, BIGR, F-75015 Paris, France; B) Université Paris-Saclay, INRAE, MetaGenoPolis, 78350, Jouy-en-Josas, France ; C) Institut de biologie de l'Ecole normale supérieure (IBENS), Ecole normale supérieure, CNRS, INSERM, PSL Universite Paris, 75005 Paris, France ; D) Centre de Recherches sur les Macromolécules Végétales, University Grenoble Alpes, CNRS,UPR 5301, Grenoble, France.

tatiana.galochkina@u-paris.fr

Protein-carbohydrate (PC) interactions govern a wide variety of biological processes and play a crucial role in the development of different diseases. During the last decades, the release of an impressive amount of data on carbohydrate-binding proteins led to the emergence of first data driven methods for prediction of carbohydrate binding sites. Nevertheless, the performance of such models remains limited as compared to similar bioinformatics problems, and its correct evaluation is hindered by the lack of the reliable and non-redundant datasets. In the current study, we address this challenge and perform an exhaustive analysis of the diversity of PC interfaces and of its impact on prediction models accuracy.

We have gathered and annotated all the available information on PC interfaces found in the Protein Data Bank (PDB) in a user-friendly web-server, DIONYSUS: https://www.dsimb.inserm.fr/DIONYSUS/. Using a customized algorithm, we identified 46,984 PC complexes interacting with one of 3,500 carbohydrate-containing ligands of the PDB (increasing the number of these structures by orders of magnitude as compared to 900 ligand names available in the Chemical Component Dictionary). We performed an exhaustive study of PC interface diversity at different levels: by functional class of interaction, protein sequence identity and local geometrical similarity between the interfaces. Furthermore, we identified representative structures of different classes of PC interactions and used them to annotate PC complexes with missing functional information.

Finally, the developed database allows us to train several deep learning models based on protein language model encoding of the protein sequence combined to molecular graphs to encode protein structure. In-depth analysis of our model performance and its comparison to the previously published methods demonstrates significant improvements of carbohydrate binding site identification as well as highlights the remaining challenges in the field.

Keywords: protein-carbohydrate interactions; protein structure; binding site analysis; machine learning; structural bioinformatics.