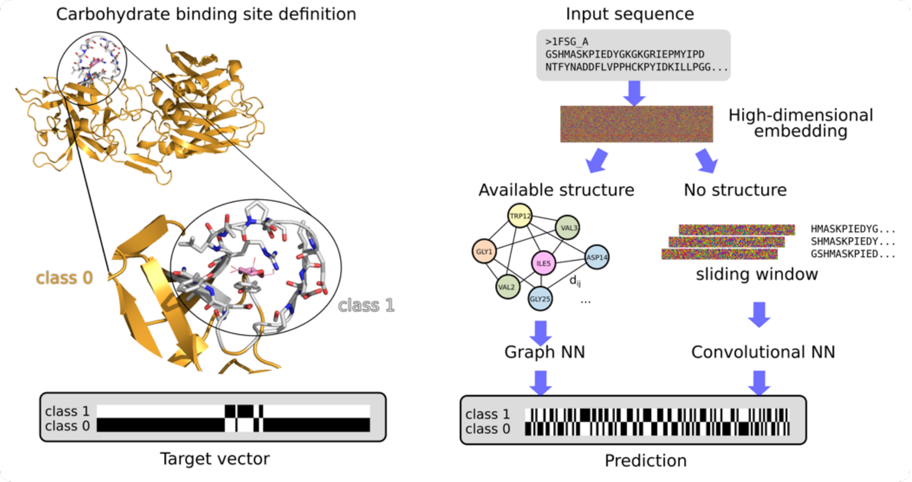
Deep learning based prediction of protein-carbohydrate interfaces

Tatiana GALOCHKINA [1,2], Aria GHEERAERT [1,2], Thomas BAILLY [1,2], Yani REN [1,2], Yann VANDER MEERSCHE [1,2], Gabriel CRETIN [1,2], Jean-Christophe GELLY [1,2]

[1] Université Paris Cité and Université des Antilles and Université de la Réunion, INSERM, BIGR, Paris, France [2] Laboratoire d'Excellence GR-Ex, Paris, France

tatiana.galochkina@u-paris.fr

Protein-carbohydrate (PC) interactions are involved in a majority of crucial biological processes. Experimental resolution of 3D structures of PC complexes is particularly difficult due to chemical and structural variability of carbohydrates and weak affinities of PC interactions, leading to underrepresentation of carbohydrates in the Protein Data Bank. Data-driven methods have potential to indicate carbohydrate binding regions proteins with missing experimental information, but are much less developed as compared to protein-protein or drug-protein interaction prediction tools. In the current study, we propose for the first time two deep learning methods for carbohydrate binding site prediction. In the first model, we use embeddings derived from the pre-trained protein language model ESM-2 [1] for efficient encoding of sequence information, where we encode each residue using sliding windows. In the second model, if the experimental structure of the protein is available, we represent proteins as amino acid contact graphs and use the positional embeddings as node features. Both models are trained to predict residues in contact with carbohydrate ligands using convolutional neural networks as architecture for the sequence-based method and graph convolutional neural network for structure-based predictions in case the structure is available. Our models outperform the existing carbohydrate-specific [2] as well as non-specific binding site prediction tools [3] and successfully detect carbohydrate binding residues missed by other methods for proteins of biological interest. Therefore, the developed methods can have an important impact for understanding mechanisms of glycan recognition as well as for carbohydrate-based drug design.



This work was supported by French National Research Agency (grant number ANR-21-CE45-0019).

Bibliographic references:

[1] Zeming Lin et al. Evolutionary-scale prediction of atomic level protein structure with a language model. bioRxiv, 2022.

[2] Masaki Banno et al. Development of a sugar binding residue prediction system from protein sequences using support vector machine. Comput. Biol. Chem., 66:36–43, 2017.

[3] Maria Littmann et al. Protein embeddings and deep learning predict binding residues for various ligand classes. Sci. Rep., 11(1):1–15, 2021